monstrous moonshine
various papers collected at a random
by a poor student

ELSEVIER

Journal of Pure and Applied Algebra 126 (1998) 233–266

JOURNAL OF PURE AND APPLIED ALGEBRA

Generalized Kac–Moody Lie algebras, free Lie algebras and the structure of the Monster Lie algebra

Elizabeth Jurisich*

Department of Mathematics, Rutgers University New Brunswick, NJ 08903, USA

Communicated by C.A. Weibel; received 4 January 1994; revised 15 May 1996

Abstract

It is shown that any generalized Kac–Moody Lie algebra $\mathfrak{g}$ that has no mutually orthogonal imaginary simple roots can be written as $\mathfrak{g} = \mathfrak{u}^+ \oplus (\mathfrak{g}_J + \mathfrak{h}) \oplus \mathfrak{u}^-$, where $\mathfrak{g}_J$ is a Kac–Moody algebra defined from a symmetrizable Cartan matrix, and $\mathfrak{u}^+$ and $\mathfrak{u}^-$ are subalgebras isomorphic to free Lie algebras over certain $\mathfrak{g}_J$-modules. The denominator identity for such an algebra $\mathfrak{g}$ is obtained by using a generalization of Witt's formula that computes the graded dimension of the free Lie algebra $\mathfrak{u}^-$ and the denominator identity known for the Kac–Moody subalgebra $\mathfrak{g}_J$. The main result and consequent proof of the denominator identity give a new proof that the radical of a generalized Kac–Moody algebra of the above type is zero. The main result is applied to the Monster Lie algebra $\mathfrak{m}$ to obtain an elegant decomposition $\mathfrak{m} = \mathfrak{u}^+ \oplus \mathfrak{gl}_2 \oplus \mathfrak{u}^-$.

Also included is a detailed discussion of Borcherds' construction of the Monster Lie algebra from a vertex algebra and an elementary proof of Borcherds' theorem relating Lie algebras with "an almost positive definite bilinear form" to generalized Kac–Moody algebras. © 1998 Elsevier Science B.V.

1991 Math. Subj Class.: 17B65; 17B67; 17B55

1. Introduction

Generalized Kac–Moody algebras, called Borcherds algebras in [14], were investigated by Borcherds in [2]. We show that any generalized Kac–Moody algebra $\mathfrak{g}$ that has no mutually orthogonal imaginary simple roots can be written as $\mathfrak{g} = \mathfrak{u}^+ \oplus (\mathfrak{g}_J + \mathfrak{h}) \oplus \mathfrak{u}^-$, where $\mathfrak{u}^+$ and $\mathfrak{u}^-$ are subalgebras isomorphic to free Lie algebras with given generators, and $\mathfrak{g}_J$ is a Kac–Moody algebra defined from a symmetrizable Cartan matrix (see

* Present address: Department of Mathematics, 5734 University Avenue, Chicago, IL 60637, USA; E-mail address: jurisich@math.uchicago.edu.

0022-4049/98/$19.00 © 1998 Elsevier Science B.V. All rights reserved
PII S0022-4049(96)00142-9

Theorem 5.1). There is a formula due to Witt that computes the graded dimension of a free Lie algebra where all of the generators have been assigned degree one. It is known that Witt's formula can be extended to other gradings (e.g., [5]). We present a further generalization of the formula appearing in [5]. The denominator identity for $\mathfrak{g}$ is obtained by using this generalization of Witt's formula and the denominator identity known for the Kac–Moody algebra $\mathfrak{g}_J$. In this work, we are taking $\mathfrak{g}$ to be the algebra defined by the appropriate generators and relations, rather than the quotient of this algebra by its radical. In particular, our main result and consequent proof of the denominator identity give a new proof that the radical of a generalized Kac–Moody algebra of the above type is zero. (We use the fact that the radical of $\mathfrak{g}_J$ is zero, which is Serre's theorem in the case that $\mathfrak{g}_J$ is finite-dimensional; this is the main case for us.)

The most important application of our work is to the Monster Lie algebra $\mathfrak{m}$, defined by Borcherds [4]. In fact, we show that $\mathfrak{m} = \mathfrak{u}^+ \oplus \mathfrak{gl}_2 \oplus \mathfrak{u}^-$, with $\mathfrak{u}^\pm$ free Lie algebras. This result is obtained by applying the above results to a generalized Kac–Moody algebra $\mathfrak{g}(M)$ defined from a particular matrix M, given by the inner products of the simple roots of $\mathfrak{m}$. Theorem 5.1 applied to this Lie algebra establishes that the subalgebras $\mathfrak{n}^\pm \subset \mathfrak{g}(M)$ are each the semidirect product of a one-dimensional Lie algebra and a free Lie algebra on countably many generators. The Lie algebra $\mathfrak{g}(M)$ is shown to be a central extension of the Monster Lie algebra $\mathfrak{m}$ (Theorem 6.1) constructed by Borcherds in [4]. By Theorem 6.1, the subalgebras $\mathfrak{m}^\pm$ in $\mathfrak{m} = \mathfrak{m}^+ \oplus \mathfrak{h} \oplus \mathfrak{m}^-$ are isomorphic to the subalgebras $\mathfrak{n}^\pm \subset \mathfrak{g}(M)$. In this way we show $\mathfrak{m}$ contains two large subalgebras $\mathfrak{u}^\pm$ which are isomorphic to free Lie algebras, and $\mathfrak{m} = \mathfrak{u}^+ \oplus \mathfrak{gl}_2 \oplus \mathfrak{u}^-$. The denominator identity for $\mathfrak{m}$ (see [4]) is obtained in this paper in the manner described above for more general $\mathfrak{g}$. In this case $\mathfrak{g}_J = \mathfrak{sl}_2$, and our results give a new proof that the central extension $\mathfrak{g}(M)$ of $\mathfrak{m}$ has zero radical.

The Monster Lie algebra $\mathfrak{m}$ is of great interest because Borcherds defines and uses this Lie algebra, along with its denominator identity, to solve the following problem [4]: It was conjectured by Conway and Norton [6] that there should be an infinite-dimensional representation of the Monster simple group such that the McKay–Thompson series of the elements of the Monster group (that is, the graded traces of the elements of the Monster group as they act on the module) are equal to some known modular functions given in [6]. After the "moonshine module" $V^\natural$ for the Monster simple group was constructed [10] and many of its properties, including the determination of some of the McKay–Thompson series, were proven in [11], the nontrivial problem of computing the rest of the McKay–Thompson series of Monster group elements acting on $V^\natural$ remained. Borcherds has shown in [4] that the McKay–Thompson series are the expected modular functions.

In this paper, in preparation for our main result, we include a detailed treatment of some of Borcherds' work on generalized Kac–Moody algebras, and of that part of [4] which shows that the Monster Lie algebra has the properties that we need to prove Theorem 6.1. We now explain this exposition.

Some results such as character formulas and a denominator identity known for Kac–Moody algebras are stated in [2] for generalized Kac–Moody algebras. We found it

necessary to do some extra work in order to understand fully the precise definitions and also the reasoning which are implicit in Borcherds' work on this subject. Kac in [18] gives an outline (without detail) of how to rigorously develop the theory of generalized Kac–Moody algebras by indicating that one should follow the arguments presented there for Kac–Moody algebras (see also [16]). Included in [17] is a detailed exposition of the theory of generalized Kac–Moody algebras, along the lines of [20], where the homology results of [13] (not covered in [18]) are extended to these new Lie algebras. That this can be done is mentioned and used in [4]. The homology result gives another proof of the character and denominator formulas [17].

We find it appropriate to work with the extended Lie algebra as in [13, 20] (that is, the Lie algebra with suitable degree derivations adjoined). Alternatively, one can generalize the theorems in [18]. In either of these approaches, the Cartan subalgebra is sufficiently enlarged to make the simple roots linearly independent and have multiplicity one, just as in the case of Kac–Moody algebras. Without working in the extended Lie algebra it does not seem possible to prove the denominator and character formulas for all generalized Kac–Moody algebras. This is because the matrix from which we define the Lie algebra can have linearly dependent columns (as in the case of $\widehat{\mathfrak{sl}_2}$); we may even have infinitely many columns equal. Naturally, when it makes sense to do so, we may specialize formulas obtained involving the root lattice. In this way we obtain Borcherds' denominator identity for $\mathfrak{m}$, and show its relation to our generalization of Witt's formula.

The crucial link between the Monster Lie algebra and a generalized Kac–Moody algebra defined from a matrix is provided by Theorem 4.1, which is a theorem given by Borcherds in [3]. Versions of this theorem also appear in [2, 4]. Since this theorem can be stated most neatly in terms of a canonical central extension of a generalized Kac–Moody algebra (as in [3]) we include a section on this central extension. Theorem 4.1 roughly says that a Lie algebra with an "almost positive definite bilinear form", like the Monster Lie algebra, is the homomorphic image of a canonical central extension of a generalized Kac–Moody algebra. The way that Theorem 4.1 is stated here and in [3] (as opposed to [4] where condition 4 is not used) allows us to conclude that the Monster Lie algebra has a central extension which is a generalized Kac–Moody algebra defined from a matrix. We include in this paper a completely elementary proof of Theorem 4.1. This proof is simpler than the argument in [2] and the proof indicated [18], which require the construction of a Casimir operator. Here Eq. (11), which follows immediately from the hypotheses of the theorem, is used in place of the Casimir operator.

The Monster Lie algebra is defined (see [4]) from the vertex algebra which is the tensor product of $V^\natural$ and a vertex algebra obtained from a rank two hyperbolic lattice. This construction is reviewed in Section 6.2. The infinite-dimensional representation $V^\natural$ of the Monster simple group constructed in [10] can be given the structure of a vertex operator algebra, as stated in [1] and proved in [11]. The theory of vertex algebras and vertex operator algebras is used in proving properties of the Monster Lie algebra, so the definition of vertex algebra and a short discussion of the properties

of vertex algebras are given in this paper. The "no-ghost" theorem of string theory is used here, as it is in [4], to obtain an isomorphism between homogeneous subspaces of the Monster Lie algebra and the homogeneous subspaces of $V^{\natural}$. A reformulation of the proof of the no-ghost theorem as given in [4, 25] is presented in the appendix of this paper.

This paper is related to the work of Kang [19], where a root multiplicity formula for generalized Kac–Moody algebras is proven by using Lie algebra homology. We recover Kang's result for the class of Lie algebras studied in this paper. Other related works include that of Harada et al. [16], who present an exposition of generalized Kac–Moody algebras along the lines of [18] (their proof of Theorem 4.1, is the proof in [2] done in complete detail; see above). The recent work of the physicists Gebert and Teschner [14] explores the module theory for some basic examples of generalized Kac–Moody algebras.

2. Generalized Kac–Moody algebras

2.1. Construction of the algebra associated to a matrix

In [2] Borcherds defines the generalized Kac–Moody algebra (GKM) associated to a matrix. Statements given here without proof have been shown in detail in [17]. In addition to [2], the reader may also want to refer to [18] where an outline is given for extending the arguments given there for Kac–Moody algebras. The Lie algebra denoted $\mathfrak{g}'(A)$ in [18], which is defined from an arbitrary matrix A, is equal to the generalized Kac–Moody algebra $\mathfrak{g}(A)$, defined below, when the matrix A satisfies conditions **(C1)**–**(C3)** given below.

Remark. In [4] any Lie algebra satisfying conditions 1–3 of Theorem 4.1 is defined to be a generalized Kac–Moody algebra. In this paper the term "generalized Kac–Moody algebra" will always mean a Lie algebra defined from a matrix as in [18]. The theory presented here, based on symmetric rather than symmetrizable matrices, can be easily adapted to the case where the matrix is symmetrizable. We use symmetric matrices in order to be consistent with the work of Borcherds and because the symmetric case is sufficient for the main applications.

We will begin by constructing a generalized Kac–Moody algebra associated to a matrix.

All vector spaces are assumed to be over $\mathbb{R}$. Let I be a set, at most countable, identified with $\mathbb{Z}_+ = \{1,2,\ldots\}$ or with $\{1,2,\ldots,k\}$. Let $A = (a_{ij})_{i,j\in I}$ be a matrix with entries in $\mathbb{R}$, satisfying the following conditions:

(C1) A is symmetric,

(C2) if $i \neq j$ then $a_{ij} \leq 0$,

(C3) if $a_{ii} > 0$ then $\frac{2a_{ij}}{a_{ii}} \in \mathbb{Z}$ for all $j \in I$.

Let $\mathfrak{g}_0(A) = \mathfrak{g}_0$ be the Lie algebra with generators h_i, e_i, f_i, where $i \in I$, and the following defining relations:

(R1) $[h_i, h_j] = 0$,

(R2) $[h_i, e_k] - a_{ik}e_k = 0$,

(R3) $[h_i, f_k] + a_{ik}f_k = 0$,

(R4) $[e_i, f_j] - \delta_{ij}h_i = 0$

for all $i, j, k \in I$. Let $\mathfrak{h} = \sum_{i\in I} \mathbb{R}h_i$. Let $\mathfrak{n}_0^+$ be the subalgebra generated by the $\{e_i\}_{i\in I}$ and let $\mathfrak{n}_0^-$ be the subalgebra generated by the $\{f_i\}_{i\in I}$. The following proposition is proven by the usual methods for Kac–Moody algebras (see [17] or [18]).

Proposition 2.1. *The Lie algebra $\mathfrak{g}_0$ has triangular decomposition $\mathfrak{g}_0 = \mathfrak{n}_0^- \oplus \mathfrak{h} \oplus \mathfrak{n}_0^+$. The abelian subalgebra $\mathfrak{h}$ has a basis consisting of $\{h_i\}_{i\in I}$, and $\mathfrak{n}_0^\pm$ is the free Lie algebra generated by the e_i (resp. the f_i) $i \in I$. In particular, $\{e_i, f_i, h_i\}_{i\in I}$ is a linearly independent set in $\mathfrak{g}_0$.*

For all $i \neq j$ and $a_{ii} > 0$ define

$$d_{ij}^+ = (\operatorname{ad} e_i)^{1-2a_{ij}/a_{ii}} e_j \in \mathfrak{g}_0^+, \tag{1}$$

$$d_{ij}^- = (\operatorname{ad} f_i)^{1-2a_{ij}/a_{ii}} f_j \in \mathfrak{g}_0^-. \tag{2}$$

Let $\mathfrak{k}_0^\pm \subset \mathfrak{n}_0^\pm$ be the ideal of $\mathfrak{n}_0^\pm$ generated by the elements:

$$d_{ij}^\pm, \tag{3}$$

$$[e_i, e_j] \quad \text{if } a_{ij} = 0 \text{ (in the case of } \mathfrak{k}_0^+), \tag{4}$$

$$[f_i, f_j] \quad \text{if } a_{ij} = 0 \text{ (in the case of } \mathfrak{k}_0^-). \tag{5}$$

Note that if $a_{ii} > 0$, then the elements (4) and (5) are of type (1) and (2). The subalgebra $\mathfrak{k}_0 = \mathfrak{k}_0^+ \oplus \mathfrak{k}_0^-$ is an ideal of $\mathfrak{g}_0$ (for details adapt the proof of Proposition 3.1 in the next section).

Definition 1. The generalized Kac–Moody algebra $\mathfrak{g}(A) = \mathfrak{g}$ associated to the matrix A is the quotient of $\mathfrak{g}_0$ by the ideal $\mathfrak{k}_0 = \mathfrak{k}_0^+ \oplus \mathfrak{k}_0^-$.

Remark. In [17, 18], a generalized Kac–Moody algebra is constructed as a quotient of $\mathfrak{g}_0$ by its radical (that is, the largest graded ideal having trivial intersection with $\mathfrak{h}$). Although this fact is not used in this paper, the ideal $\mathfrak{k}_0$ is equal to the radical of $\mathfrak{g}_0$. Of course, proving that the radical of the Lie algebra $\mathfrak{g}(A)$ is zero not trivial. It is shown in [17] that the radical of $\mathfrak{g}(A)$ is zero using results from [12, 18] and a proposition proven in [2].

Let $\mathfrak{n}^+ = \mathfrak{n}_0^+/\mathfrak{k}_0^+$ and $\mathfrak{n}^- = \mathfrak{n}_0^-/\mathfrak{k}_0^-$. Proposition 2.1 implies that the generalized Kac–Moody algebra has triangular decomposition $\mathfrak{g} = \mathfrak{n}^+ \oplus \mathfrak{h} \oplus \mathfrak{n}^-$. The Lie algebra $\mathfrak{g}$ is given by the corresponding generators and relations. The Lie algebra $\mathfrak{g}(A)$ is a Kac–Moody algebra when the matrix A is a generalized Cartan matrix.

Let $\deg e_i = -\deg f_i = (0,\ldots,0,1,0,\ldots)$ where 1 appears in the ith position, and let $\deg h_i = (0,\ldots)$. This induces a Lie algebra grading by $\mathbb{Z}^I$ on $\mathfrak{g}$. Degree derivations D_i are defined by letting D_i act on the degree $(n_1,n_2,\ldots)$ subspace of $\mathfrak{g}$ as multiplication by the scalar n_i. Let $\mathfrak{d}$ be the space spanned by the D_i. We extend the Lie algebra $\mathfrak{g}$ by taking the semidirect product with $\mathfrak{d}$, so $\mathfrak{g}^e = \mathfrak{d} \ltimes \mathfrak{g}$. Then $\mathfrak{h}^e = \mathfrak{d} \oplus \mathfrak{h}$ is an abelian subalgebra of $\mathfrak{g}^e$, which acts via scalar multiplication on each space $\mathfrak{g}(n_1,n_2,\ldots)$.

Let $\alpha_i \in (\mathfrak{h}^e)^*$ for $i \in I$ be defined by the conditions

$$[h,e_i] = \alpha_i(h)e_i \quad \text{for all } h \in \mathfrak{h}^e.$$

Note that $\alpha_j(h_i) = a_{ij}$ for all $i,j \in I$. Because we have adjoined $\mathfrak{d}$ to $\mathfrak{h}$, the α_i are linearly independent.

For all $\varphi \in (\mathfrak{h}^e)^*$ define

$$\mathfrak{g}^\varphi = \{x \in \mathfrak{g} \mid [h,x] = \varphi(h)x \ \forall h \in \mathfrak{h}^e\}.$$

If $\varphi,\psi \in (\mathfrak{h}^e)^*$ then $[\mathfrak{g}^\varphi,\mathfrak{g}^\psi] \subset \mathfrak{g}^{\varphi+\psi}$. By definition $e_i \in \mathfrak{g}^{\alpha_i}$, and $f_i \in \mathfrak{g}^{-\alpha_i}$ for all $i \in I$. If all $n_i \leq 0$, or all $n_i \geq 0$ (only finitely many nonzero), it can be shown by using the same methods as for Kac–Moody algebras that

$$\mathfrak{g}^{n_1\alpha_1+n_2\alpha_2+\cdots} = \mathfrak{g}(n_1,n_2,\ldots)$$

and $\mathfrak{g}^0 = \mathfrak{h}$. Therefore,

$$\mathfrak{g} = \coprod_{\substack{(n_1,n_2,\ldots)\\ n_i\in\mathbb{Z}}} \mathfrak{g}^{n_1\alpha_1+n_2\alpha_2+\cdots}. \tag{6}$$

Definition 2. The *roots* of $\mathfrak{g}$ are the nonzero elements φ of $(\mathfrak{h}^e)^*$ such that $\mathfrak{g}^\varphi \neq 0$. The elements α_i are *simple roots*, and $\mathfrak{g}^\varphi$ is the *root space* of $\varphi \in (\mathfrak{h}^e)^*$.

Denote by Δ the set of roots, Δ_+ the set of *positive* roots, i.e., the non-negative integral linear combinations of α_i. Let $\Delta_- = -\Delta_+$ be the set of *negative* roots. All of the roots are either positive or negative.

The algebra $\mathfrak{g}$ has an automorphism η of order 2 which acts as -1 on $\mathfrak{h}$ and interchanges the elements e_i and f_i. By an inductive argument, as in [18] or [23], we can construct a symmetric invariant bilinear form on $\mathfrak{g}$ such that g^φ and $\mathfrak{g}^{-\varphi}$ where $\varphi \in \Delta_+$ are nondegenerately paired; however, the restriction of this form to $\mathfrak{h}$ can be degenerate. There is a character formula for standard modules of $\mathfrak{g}$ and a denominator identity (see [2, 17, 18])

$$\prod_{\varphi\in\Delta_+}(1-e^\varphi)^{\dim g^\varphi} = \sum_{w\in W}(\det w)\sum_{\gamma\in\Omega(0)}(-1)^{l(\gamma)}e^{w(\rho+\gamma)-\rho}, \tag{7}$$

where $\Omega(0) \subset \Delta_+ \cup \{0\}$ is the set of all $\gamma \in \Delta_+ \cup \{0\}$ such that γ is the sum (of length zero or greater) of mutually orthogonal imaginary simple roots.

Remark. The denominator formula (7) can be specialized to the unextended Lie algebra as long as the resulting specialization is well defined.

3. A canonical central extension

It is useful to consider a certain central extension of the generalized Kac–Moody algebra. Working with the central extension defined here (which is the same as in [3]) will simplify the statement and facilitate the proof of Theorem 4.1 below. Given a matrix A satisfying **(C1)**–**(C3)** let $\hat{\mathfrak{g}}$ be the Lie algebra with generators e_i, f_i, h_{ij} for $i, j, k, l \in I$ and relations:

(R1′) $[h_{ij}, h_{kl}] = 0$,

(R2′) $[h_{ij}, e_k] - \delta_{i,j} a_{ik} e_k = 0$ and $[h_{ij}, f_k] + \delta_{i,j} a_{ik} f_k = 0$,

(R3′) $[e_i, f_j] - h_{ij} = 0$,

(R4′) $d_{ij}^{\pm} = 0$ for all $i \neq j, a_{ii} > 0$,

(R5′) If $a_{ij} = 0$ then $[e_i, e_j] = 0$ and $[f_i, f_j] = 0$.

The elements $d_{ij}^{\pm}$ are defined by (1) and (2).

We will study this Lie algebra by first considering the Lie algebra $\hat{\mathfrak{g}}_0$ with generators h_{ij}, e_i, f_i, where $i, j \in I$, and the defining relations **(R1′)**–**(R3′)**.

Let $\hat{\mathfrak{h}} = \sum_{i,j \in I} \mathbb{R} h_{ij}$ and let $\hat{\mathfrak{n}}_0^+$ be the subalgebra generated by the e_i $i \in I$, $\hat{\mathfrak{n}}_0^-$ be the subalgebra generated by the f_i $i \in I$. We shall prove a version of Proposition 2.1.

Lemma 3.1. *The elements h_{ij} are zero unless the ith and jth columns of the matrix A are equal.*

Proof (*Also see Borcherds* [3]). The lemma follows from the Jacobi identity and relations **(R1′)**, **(R2′)**. □

Proposition 3.1. *The Lie algebra $\hat{\mathfrak{g}}_0 = \hat{\mathfrak{n}}_0^- \oplus \hat{\mathfrak{h}} \oplus \hat{\mathfrak{n}}_0^+$, and the abelian Lie algebra $\hat{\mathfrak{h}}$ has a basis consisting of $\{h_{ij}\}_{i,j \in I}$ such that the* ith *and* jth *columns of $A = (a_{ij})_{i,j \in I}$ are equal. The subalgebra $\hat{\mathfrak{n}}_0^{\pm}$ is the free Lie algebra generated by the e_i (resp. the f_i), $i \in I$. The set $\{e_i, f_i\}_{i,j \in I} \cup \{h_{ij}\}_{i,j \in S}$ is linearly independent where $S = \{(i,j) \in I \times I \mid a_{ki} = a_{kj}$ for all $k \in I\}$.*

Proof. As in the classical case, one constructs a sufficiently large representation of the Lie algebra. Let $\mathfrak{h}$ be the span of the elements $\{h_{ij}\}_{i,j \in I}$. Define $\alpha_j \in \mathfrak{h}^*$ as follows:

$$\alpha_j(h_{ik}) = \delta_{ik} a_{ij}.$$

Let X be the free associative algebra on the symbols $\{x_i\}_{i \in I}$. Let $\lambda \in \mathfrak{h}^*$ be such that $\lambda(h_{ij}) = 0$ if $a_{li} \neq a_{lj}$, i.e., unless the ith and jth columns of A are equal.

We define a representation of the free Lie algebra $\mathfrak{g}_F$ with generators e_i, f_i, h_{ij} on X by the following actions of the generators:

1. $h \cdot 1 = \lambda(h)$ for all $h \in \mathfrak{h}$,
2. $f_i \cdot 1 = x_i$ for all $i \in I$,
3. $e_i \cdot 1 = 0$ for all $i \in I$,
4. $h \cdot x_{i_1} \cdots x_{i_r} = (\lambda - \alpha_{i_1} - \cdots - \alpha_{i_r})(h) x_{i_1} \cdots x_{i_r}$ for $h \in \mathfrak{h}$,

5. $f_i \cdot x_{i_1} x_{i_2} \cdots x_{i_r} = x_i x_{i_1} x_{i_2} \cdots x_{i_r}$,

6. $e_i \cdot x_{i_1} \cdots x_{i_r} = x_{i_1} e_i \cdot x_{i_2} \cdots x_{i_r} + (\lambda - \alpha_{i_2} - \cdots - \alpha_{i_r})(h_{ii_1}) x_{i_2} x_{i_3} \cdots x_{i_r}$.

Let $\hat{\mathfrak{s}}$ be the ideal generated by the elements $[h_{ij}, h_{kl}]$, $[h_{ij}, e_k] - \delta_{i,j} a_{ik} e_k$, $[h_{ij}, f_k] + \delta_{i,j} a_{ik} f_k$, and $[e_i, f_j] - h_{ij}$. Now we will show that the ideal $\hat{\mathfrak{s}}$ annihilates the module X: It is clear that $[h, h']$ is 0 on X for $h, h' \in \mathfrak{h}$. The element $[e_i, f_i] - h_{ij}$ also acts as 0 on X by the following computation:

$$\begin{aligned}
[e_i, f_j] \cdot x_{i_1} \cdots x_{i_r} &= e_i f_j \cdot x_{i_1} \cdots x_{i_r} - f_j e_i \cdot x_{i_1} \cdots x_{i_r} \\
&= e_i \cdot x_j x_{i_1} \cdots x_{i_r} - x_j e_i \cdot x_{i_1} \cdots x_{i_r} \\
&= (\lambda - \alpha_{i_1} - \cdots - \alpha_{i_r})(h_{ij}) x_{i_1} \cdots x_{i_r} \\
&= h_{ij} \cdot x_{i_1} \cdots x_{i_r}.
\end{aligned}$$

By a similar computation $[h_{ij}, f_k] + \delta_{ij} a_{ik} f_k$ annihilates X. Now consider the action of $[h_{ij}, e_k] - \delta_{ij} a_{ik} e_k$ on X:

$$\begin{aligned}
[h_{ij}, e_k] \cdot 1 &= h_{ij} e_k \cdot 1 - e_k h_{ij} \cdot 1 \\
&= e_k \cdot \lambda(h_{ij}) 1 \\
&= 0
\end{aligned}$$

and

$$\delta_{ij} a_{ik} e_k \cdot 1 = 0.$$

Thus, $[h_{ij}, e_k] - \delta_{ij} a_{ik} e_k$ annihilates 1. Furthermore, $[h_{ij}, e_k] - \delta_{ij} a_{ik} e_k$ commutes with the action of f_l for all l, as the following computation from [3] shows

$$\begin{aligned}
&[[h_{ij}, e_k] - \delta_{ij} a_{ik} e_k, f_l] \\
&\quad = [[h_{ij}, e_k], f_l\ ] - [\delta_{ij} a_{ik} e_k, f_l] = [h_{ij}, h_{kl}] + \delta_{ij}(a_{il} - a_{ik}) h_{kl}.
\end{aligned}$$

By the assumption on λ, h_{kl} is 0 on X unless $a_{il} = a_{ik}$, so that the above is zero. Since any $x_{i_1} \cdots x_{i_r} = f_{i_1} \cdots f_{i_r} \cdot 1$, this means that $[h_{ij}, e_k] - \delta_{ij} a_{ik} e_k$ annihilates X. Now X can be regarded as a $\hat{\mathfrak{g}}_0$-module. The remainder of the proof follows the classical argument. □

The following proposition will be used in the proof of Theorem 4.1.

Proposition 3.2. *In $\hat{\mathfrak{g}}_0$, for all $i, j, k \in I$ with $i \neq j$ and $a_{ii} > 0$*

$$[e_k, d_{ij}^-] = 0$$

and

$$[f_k, d_{ij}^+] = 0.$$

Proof. It is enough to show the first formula.

Case 1: Assume $k \neq i$ and $k \neq j$. Since h_{ki} is central if $k \neq i$

$$\begin{aligned}(\operatorname{ad} e_k)(\operatorname{ad} f_i)x &= [h_{ki},x] + [f_i,[e_k,x]] \\ &= (\operatorname{ad} f_i)(\operatorname{ad} e_k)x,\end{aligned} \tag{8}$$

so

$$\begin{aligned}[e_k,(\operatorname{ad} f_i)^{-2a_{ij}/a_{ii}+1} f_j] &= (\operatorname{ad} f_i)^{-2a_{ij}/a_{ii}+1}[e_k,f_j] \\ &= (\operatorname{ad} f_i)^{-2a_{ij}/a_{ii}+1} h_{kj} = 0.\end{aligned}$$

The last equality holds because $k \neq j$ means h_{kj} is central.

Case 2: Assume $k = i$. By assumption $a_{ii} > 0$, thus e_i, f_i, h_{ii} generate a Lie algebra isomorphic to $\mathfrak{sl}_2$. Consider the $\mathfrak{sl}_2$-module generated by the weight vector f_j. Then if $a_{ij} = 0$ the result follows from the Jacobi identity and the fact that $[e_i, f_j] = h_{ij}$ is in the center of the Lie algebra. If $a_{ij} \neq 0$ then

$$\begin{aligned}&\operatorname{ad} e_i(\operatorname{ad} f_i)^{-2a_{ij}/a_{ii}+1} f_j \\ &\quad = (a_{ii}/2)(2a_{ij}/a_{ii} - 1)(-2a_{ij}/a_{ii})(\operatorname{ad} f_i)^{-2a_{ij}/a_{ii}} f_j \\ &\qquad +(-2a_{ij}/a_{ii} + 1)(\operatorname{ad} f_i)^{-2a_{ij}/a_{ii}}(\operatorname{ad} h_{ii}) f_j \\ &\quad = 0.\end{aligned}$$

Case 3: Assume $k = j$. By (8)

$$[e_j,(\operatorname{ad} f_i)^{-2a_{ij}/a_{ii}+1} f_j] = (\operatorname{ad} f_i)^{-2a_{ij}/a_{ii}+1} h_{jj}. \tag{9}$$

Since $[f_i, h_{jj}] = a_{ji} f_i$ it follows immediately that (9) equals zero if $a_{ij} \leq 0$. □

Let $\mathfrak{k}_0^{\pm}$ be the ideal of $\hat{\mathfrak{g}}_0^{\pm}$, respectively, generated by the elements which give relations (**R4′**) and (**R5′**):

$d_{ij}^{\pm}$ for all $i \neq j$, $a_{ii} > 0$,

$[e_i, e_j]$ if $a_{ij} = 0$ for $\mathfrak{k}_0^{+}$,

$[f_i, f_j]$ if $a_{ij} = 0$ for $\mathfrak{k}_0^{-}$.

Proposition 3.3. *Define* $\mathfrak{k}_0 = \mathfrak{k}_0^{+} \oplus \mathfrak{k}_0^{-}$. *Then* $\mathfrak{k}_0^{\pm}$ *and* $\mathfrak{k}_0$ *are ideals of* $\hat{\mathfrak{g}}_0$.

Proof. Let $\hat{\mathfrak{g}}_0$ act on itself by the adjoint representation. To see that $\mathfrak{k}^{-}$ is an ideal, first consider the action on the generators d_{ij}^{-}. By Proposition 3.1 we have

$$\begin{aligned}\sum_{i\neq j} \mathcal{U}(\hat{\mathfrak{g}}_0)\cdot d_{ij}^{-} &= \sum_{i\neq j} \mathcal{U}(\hat{\mathfrak{g}}_0^{-})\mathcal{U}(\hat{\mathfrak{h}})\mathcal{U}(\hat{\mathfrak{g}}_0^{+})\cdot d_{ij}^{-} \\ &= \sum_{i\neq j} \mathcal{U}(\hat{\mathfrak{g}}_0^{-})\cdot d_{ij}^{-} \subset \mathfrak{k}_0^{-}.\end{aligned}$$

The equality holds by Proposition 3.2 and the fact that $h(\operatorname{ad} f_i)^N f_j = \lambda(\operatorname{ad} f_i)^N f_j$, where λ is a scalar.

We must also consider the action on the generators of the form $[f_i, f_j]$, $i \neq j$, $a_{ij} = 0$. In this case

$$\begin{aligned}[e_k[f_i, f_j]] &= [h_{ki}, f_j] + [f_i, h_{kj}] \\ &= -\delta_{ki} a_{ij} f_j + \delta_{kj} a_{ji} f_i \\ &= 0,\end{aligned}$$

i.e.,

$$e_k \cdot [f_i, f_j] = 0.$$

Then by the same argument as above

$$\sum_{i \neq j,\ a_{ij}=0} \mathscr{U}(\hat{\mathfrak{g}}_0) \cdot [f_i, f_j] \subset \mathfrak{k}_0^-.$$

By a symmetric argument, $\mathfrak{k}_0^+$ is an ideal of $\hat{\mathfrak{g}}_0$, so $\mathfrak{k}_0$ is an ideal. □

The Lie algebra $\hat{\mathfrak{g}}_0/\mathfrak{k}_0$ is equal to $\hat{\mathfrak{g}}$, the Lie algebra defined above by generators and relations.

Remark. The Lie algebra $\hat{\mathfrak{g}}$ is called the universal generalized Kac–Moody algebra in [3].

Let $\mathfrak{c}$ be the ideal of $\hat{\mathfrak{g}}$ spanned by the h_{ij} where $i \neq j$; note that these elements are central. The Lie algebra $\hat{\mathfrak{g}}$ is a central extension of the Lie algebra $\mathfrak{g}$, because there is an obvious homomorphism from $\hat{\mathfrak{g}}$ to $\mathfrak{g}$ mapping generators to generators with kernel $\mathfrak{c}$. So $1 \to \mathfrak{c} \to \hat{\mathfrak{g}} \to \mathfrak{g} \to 1$ is exact. The radical of $\hat{\mathfrak{g}}$ must be zero because the radical of $\mathfrak{g}$ is zero. We have shown the following.

Theorem 3.1. *The generalized Kac–Moody algebra* $\mathfrak{g}$ *is isomorphic to* $\hat{\mathfrak{g}}/\mathfrak{c}$.

The Lie algebra $\hat{\mathfrak{g}}$ can be given a $\mathbb{Z}$-gradation defined by taking $\deg e_i = -\deg f_i = s_i \in \mathbb{Z}$ where $s_i = s_j$ if $h_{ij} \neq 0$, and $\deg h_{ij} = 0$. The automorphism η is well defined on $\hat{\mathfrak{g}}$. It follows from Propositions 2.1 and 3.1 that $\mathfrak{n}_0^{\pm} = \hat{\mathfrak{n}}_0^{\pm}$. If $\mathfrak{n}^+ = \hat{\mathfrak{g}}_0{}^+/\mathfrak{k}_0^+$ and $\mathfrak{n}^- = \hat{\mathfrak{g}}_0{}^-/\mathfrak{k}_0^-$ then we have the decomposition $\hat{\mathfrak{g}} = \mathfrak{n}^- \oplus \hat{\mathfrak{h}} \oplus \mathfrak{n}^+$.

Recall that the Lie algebra $\mathfrak{g}^e$ has an invariant bilinear form $(\cdot,\cdot)$. It is useful to define an invariant bilinear form $(\cdot,\cdot)_{\hat{\mathfrak{g}}}$ on $\hat{\mathfrak{g}}$. For $a, b \in \hat{\mathfrak{g}}$, let $(a,b)_{\hat{\mathfrak{g}}} = (\bar{a}, \bar{b})$. Note that the span of the h_{ij} where $i \neq j$ is in the radical of the form on $\hat{\mathfrak{g}}$. The form $(\cdot,\cdot)_{\hat{\mathfrak{g}}}$ is symmetric and invariant because the form on $\mathfrak{g}$ has these properties. Grade $\hat{\mathfrak{g}}$ by letting $\deg e_i = 1 = -\deg f_i$, and $\deg h_{ij} = 0$, so $\hat{\mathfrak{g}} = \bigoplus_{n \in \mathbb{Z}} \hat{\mathfrak{g}}_n$, where $\hat{\mathfrak{g}}_n$ is contained in $\mathfrak{n}^+$ if $n > 0$, and $\mathfrak{n}^-$ if $n < 0$. The form $(\cdot,\cdot)_{\hat{\mathfrak{g}}}$ is nondegenerate on $\hat{\mathfrak{g}}_n \oplus \hat{\mathfrak{g}}_{-n}$, because the

map given by the central extension, $x \mapsto \bar{x}$, is an isomorphism on $\mathfrak{n}^{\pm}$, and the form defined on $\mathfrak{g}^e$ is nondegenerate on $\mathfrak{g}^e_m \oplus \mathfrak{g}^e_{-m}$ for $m \in \mathbb{Z}_+$.

4. Another characterization of GKM algebras

Theorem 4.1 below is a version of Theorem 3.1 appearing in [2]. Much of the proof of the following theorem is different than the proof appearing in [2]. In particular, there is no need to define a Casimir operator in order to show that the elements $a_{ij} \leq 0$; as seen below, this follows immediately from condition 3.

Remark. In [4] Borcherds states, as a converse to the theorem, that the canonical central extensions $\hat{\mathfrak{g}}$ satisfy conditions 1–3 below, although we note that the canonical central extension of a generalized Kac–Moody algebra does not have to satisfy condition 1 for some matrices. For example, if we start with the infinite matrix whose entries are all -2, then all of the e_i must have the same degree because of condition 3. Therefore, there is no way to define a $\mathbb{Z}$-grading of $\hat{\mathfrak{g}}$ so that $\hat{\mathfrak{g}}_i$ is both finite-dimensional and satisfies condition 3.

We also note that the kernel of the map π appearing in the following theorem can be strictly larger that the span of the h_{ij}; cf. the statement of Theorem 4.1 in [4].

Theorem 4.1 (Borcherds [4]). *Let* $\mathfrak{g}$ *be a Lie algebra satisfying the following conditions*:

1. $\mathfrak{g}$ *can be* $\mathbb{Z}$*-graded as* $\coprod_{i\in\mathbb{Z}} \mathfrak{g}_i$, $\mathfrak{g}_i$ *is finite dimensional if* $i \neq 0$, *and* $\mathfrak{g}$ *is diagonalizable with respect to* $\mathfrak{g}_0$.
2. $\mathfrak{g}$ *has an involution* ω *which maps* $\mathfrak{g}_i$ *onto* $\mathfrak{g}_{-i}$ *and acts as* -1 *on noncentral elements of* $\mathfrak{g}_0$, *in particular* $\mathfrak{g}_0$ *is abelian.*
3. $\mathfrak{g}$ *has a Lie algebra-invariant bilinear form* $(\cdot,\cdot)$, *invariant under* ω, *such that* $\mathfrak{g}_i$ *and* $\mathfrak{g}_j$ *are orthogonal if* $i \neq -j$, *and such that the form* $(\cdot,\cdot)_0$, *defined by* $(x,y)_0 = -(x,\omega(y))$ *for* $x, y \in \mathfrak{g}$, *is positive definite on* $\mathfrak{g}_m$ *if* $m \neq 0$.
4. $\mathfrak{g}_0 \subset [\mathfrak{g},\mathfrak{g}]$.

Then there is a central extension $\hat{\mathfrak{g}}$ *of a generalized Kac–Moody algebra and a homomorphism,* π, *from* $\hat{\mathfrak{g}}$ *onto* $\mathfrak{g}$, *such that the kernel of* π *is in the center of* $\hat{\mathfrak{g}}$.

Proof. Generators of the Lie algebra $\mathfrak{g}$ are constructed, as in [2], as follows: For $m > 0$, let $\mathfrak{l}_m$ be the subalgebra of $\mathfrak{g}$ generated by the $\mathfrak{g}_n$ for $0 < n < m$, and let $\mathfrak{e}_m$ be the orthogonal complement of $\mathfrak{l}_m$ in $\mathfrak{g}_m$ under $(\cdot,\cdot)_0$. To see that $\mathfrak{e}_m$ is invariant under $\mathfrak{g}_0$ let $x \in \mathfrak{g}_0$, $y \in \mathfrak{e}_m$, and $z \in \mathfrak{l}_m$. Then $[x,z] \in \mathfrak{l}_m$, so $(y,[x,z])_0 = 0$. Since the form $(\cdot,\cdot)_0$ satisfies $([x,y],z)_0 = -(y,[\omega(x),z])_0$, i.e., is contravariant, we have $([x,y],z)_0 = 0$, which implies that $[x,y] \in \mathfrak{e}_m$. The operators induced by the action of $\mathfrak{g}_0$ on $\mathfrak{e}_m$ commute, so we can construct a basis of $\mathfrak{e}_m$ consisting of weight vectors with respect to $\mathfrak{g}_0$. The form $(\cdot,\cdot)_0$ is positive definite on $\mathfrak{e}_m$ so an orthonormal basis can be constructed. Contravariance of the form ensures that this orthonormal basis also

consists of weight vectors. The union of these bases for all the $\mathfrak{e}_m$'s can be indexed by $I = \mathbb{Z}_+$, in any order, and will be denoted $\{e_i\}_{i\in I}$. Each $\mathfrak{g}_n$, $n > 0$, is in the Lie algebra generated by $\{e_i\}_{i\in I}$, as is seen by the following induction on the degree, n: For $n = 1$, $\mathfrak{g}_1 = \mathfrak{e}_1$. Now assume that the $\mathfrak{g}_n$ for all $0 < n < m$ are contained in the Lie algebra generated by the e_i, $i \in I$. The finite-dimensional space $\mathfrak{g}_m$ decomposes under $(\cdot,\cdot)_0$ as $\mathfrak{g}_m = \mathfrak{e}_m \oplus \mathfrak{e}_m^{\perp}$, where $\mathfrak{e}_m^{\perp} = \mathfrak{l}_m \cap \mathfrak{g}_m$. By the induction assumption, $\mathfrak{e}_m^{\perp}$ is generated by some of the e_i's, and by construction $\mathfrak{e}_m$ has a basis consisting of e_i's. Define $f_i = -\omega(e_i)$, and $h_{ij} = [e_i, f_j]$. The $\mathfrak{g}_n$ where $n < 0$ are generated by the f_i, $i \in I$.

The elements h_{ij} can be nonzero only when $\deg e_i = \deg e_j$. This is because if $\deg e_i > \deg e_j$, which can be assumed without loss of generality, then $[e_j,[e_i,f_j]] \in \mathfrak{l}_{\deg e_i}$. Thus $([e_i,f_j],[e_i,f_j])_0 = 0$ by contravariance, and $[e_i,f_j] = 0$ by the positive definiteness of $(\cdot,\cdot)_0$. Therefore, all of the h_{ij} are in $\mathfrak{g}_0$. By assumption 4, $\mathfrak{g}_0$ is generated by the h_{ij}, $i,j \in I$.

Let $k \in \mathrm{rad}(\cdot,\cdot)$ (so $k \in \mathfrak{g}_0$). Then $([k,g],[k,g])_0 = (k,[\omega(g),[k,g]])_0 = 0$ for all $g \in \mathfrak{g}_n$, $n \neq 0$. Thus, $[k,g] = 0$ by positive definiteness, and since $k \in \mathfrak{g}_0$, k must also commute with $\mathfrak{g}_0$. Therefore, the radical of the form $(\cdot,\cdot)$ is in the center of $\mathfrak{g}$.

If $i \neq j$ then $[e_i,f_j] = h_{ij}$ is contained in the center of $\mathfrak{g}$ because it is in the radical of the form $(\cdot,\cdot)$. To see this, consider $([e_i,f_j],x)$ for a homogeneous $x \in \mathfrak{g}$. We know by assumption that this is zero if $x \notin \mathfrak{g}_0$. If $x \in \mathfrak{g}_0$ then $([e_i,f_j],x) = (e_i,[f_j,x]) = c(e_i,e_j)_0 = 0$ for some real number c, as f_j is a weight vector for $\mathfrak{g}_0$, and the e_k's are orthogonal. Thus, $([e_i,f_j],x) = 0$ for all $x \in \mathfrak{g}$.

Now it must be shown that the generators constructed above satisfy the relations (**R1′**)–(**R5′**) of the central extension of a generalized Kac–Moody algebra, and the symmetric matrix with entries $(h_{ii},h_{jj}) = a_{ij}$ satisfies conditions (**C2**) and (**C3**).

That (**R1′**) (that is, $[h_{ij},h_{kl}] = 0$), holds is obvious because $\mathfrak{g}_0$ is abelian, and (**R3′**) is true by definition.

The relations (**R2′**) are proven by the following argument (cf. [2]): By construction, e_i is a weight vector of $\mathfrak{g}_0$, thus for some real number c, $[h_{lm},e_i] = ce_i$ and

$$\begin{aligned} c = c(e_i,e_i)_0 &= ([h_{lm},e_i],e_i)_0 \\ &= ([h_{lm},e_i],f_i) \\ &= (h_{lm},[e_i,f_i]) = (h_{lm},h_{ii}). \end{aligned}$$

Since h_{ij} for $i \neq j$ is in the radical of the form $(\cdot,\cdot)$, we have $c = \delta_{lm}a_{li}$. Applying ω shows the relation for f_i.

To show (**R5′**) and (**C2**) let $i \neq j$. We must prove $a_{ij} \leq 0$, and if $a_{ij} = 0$ then $[e_i,e_j] = 0$ and $[f_i,f_j] = 0$. Consider the element $[e_i,e_j] \in \mathfrak{g}$, we have

$$\begin{aligned} ([e_i,e_j],[e_i,e_j])_0 &= -(e_i,[f_j,[e_i,e_j]])_0 \\ &= -a_{ij}(e_i,e_i)_0. \end{aligned} \tag{10}$$

The last equality follows from

$$[f_j,[e_i,e_j]] = -[e_j,[f_j,e_i]] - [e_i,[e_j,f_j]]$$
$$= a_{ij}e_i.$$

By Eq. (10) and the positive definiteness of the form $(\cdot,\cdot)_0$ on $\mathfrak{g}_m$, $m \neq 0$, we have $a_{ij} \leq 0$, and $a_{ij} = 0$ if and only if $[e_i,e_j] = 0$. By applying ω we also show $[f_i,f_j] = 0$ in this case.

For $a_{ii} > 0$ the Lie algebra generated by e_i, f_i, h_{ii} is isomorphic to $\mathfrak{sl}_2$, rescaling e_i and h_{ii} by $2/a_{ii}$. For each $j \in I$ the element f_j generates an $\mathfrak{sl}_2$-weight module, the weights must all be integers so $[h_{ii},f_j] = (-2a_{ij}/a_{ii})f_j$ implies that $2a_{ij}/a_{ii} \in \mathbb{Z}$, thus **(C3)** is satisfied. Proposition 3.2, which holds for $\mathfrak{g}$, shows that $[e_k, d_{ij}^-] = 0$ where $d_{ij}^- = (\operatorname{ad} f_i)^{n+1} f_j$ and $n = -2a_{ij}/a_{ii}$ (note that n is positive). Contravariance of the form gives us

$$((\operatorname{ad} f_i)^{n+1} f_j, (\operatorname{ad} f_i)^{n+1} f_j)_0 = ((\operatorname{ad} f_i)^n f_j, [e_i, (\operatorname{ad} f_i)^{n+1} f_j])_0 = 0,$$

so that $(\operatorname{ad} f_i)^{1-2a_{ij}/a_{ii}} f_j = 0$ by the positive definiteness of $(\cdot,\cdot)_0$. Applying ω to d_{ij}^- gives the relation for $(\operatorname{ad} e_i)^{1-2a_{ij}/a_{ii}} e_j$. This shows the relations **(R4′)**.

Denote by $\hat{\mathfrak{g}}$ the canonical central extension, defined in Section 3, of the generalized Kac–Moody algebra associated to the matrix $a_{ij} = (h_{ii},h_{jj})$. Define a homomorphism $\pi : \hat{\mathfrak{g}} \to \mathfrak{g}$, taking the generators e_i, f_i, h_{ij} in $\hat{\mathfrak{g}}$ to the generators e_i, f_i, h_{ij} in $\mathfrak{g}$. Since the generators of $\mathfrak{g}$ have been shown to satisfy the relations of $\hat{\mathfrak{g}}$ the map π is a homomorphism from $\hat{\mathfrak{g}}$ onto $\mathfrak{g}$. The bilinear form, $(\cdot,\cdot)_{\hat{\mathfrak{g}}}$, on $\hat{\mathfrak{g}}$ satisfies

$$(e_i,f_j)_{\hat{\mathfrak{g}}} = \delta_{ij} = (e_i,f_j),$$

$$(h_{ii},h_{jj})_{\hat{\mathfrak{g}}} = a_{ij} = (h_{ii},h_{jj}).$$

The h_{ij} with $i \neq j$ are in the radical of $(\cdot,\cdot)_{\hat{\mathfrak{g}}}$. Thus, $(x,y)_{\hat{\mathfrak{g}}} = (\pi(x),\pi(y))$ for $x,y \in \hat{\mathfrak{g}}$, because $(x,y)_{\hat{\mathfrak{g}}}$ can be reduced using invariance and the Jacobi identity to some polynomial in the (e_i,f_j) and (h_{ii},h_{jj}).

Now we determine the kernel of the map π. Let $a \in \hat{\mathfrak{g}}$ such that $a \neq 0$ and $\pi(a) = 0$ in $\mathfrak{g}$. Recall the grading $\hat{\mathfrak{g}} = \bigoplus_{n\in\mathbb{Z}} \hat{\mathfrak{g}}_n$ and the decomposition $\hat{\mathfrak{g}} = n^+ \oplus \hat{\mathfrak{h}} \oplus n^-$, thus can write $a = a_+ + a_0 + a_-$ where $a_\pm \in n^\pm, a_0 \in \hat{\mathfrak{h}}$. Thus, $\pi(a) = \pi(a_+) + \pi(a_0) + \pi(a_-) = 0$, which is still a direct sum in $\mathfrak{g}$. Therefore, $\pi(a_+) = 0, \pi(a_0) = 0, \pi(a_-) = 0$. Assume that a_+ is homogeneous and nonzero, then for some $n > 0$, $a_+ \in \hat{\mathfrak{g}}_n$ and $(a_+,x)_{\hat{\mathfrak{g}}} = (\pi(a_+),\pi(x)) = 0$ for all $x \in \hat{\mathfrak{g}}_{-n}$. Since $(\cdot,\cdot)_{\hat{\mathfrak{g}}}$ is nondegenerate on $\hat{\mathfrak{g}}_n \oplus \hat{\mathfrak{g}}_{-n}$ we have $a_+ = 0$. By a similar argument $a_- = 0$. Thus, $a = a_0 \in \hat{\mathfrak{h}}$, and $[a,h] = 0$ for all $h \in \hat{\mathfrak{h}}$. Since $\pi(a) = 0$, we have

$$(a,h_{ii})_{\hat{\mathfrak{g}}} = (\pi(a),\pi(h_{ii})) = 0 \quad \text{for all } i$$

so

$$[a,e_i] = (a,h_{ii})_{\hat{\mathfrak{g}}} e_i = 0 \quad \text{for all } i$$

similarly $[a,f_i] = 0$ for all $i \in I$. Thus, a is in the center of $\hat{\mathfrak{g}}$. □

If the radical of the form $(\cdot,\cdot)$ is zero then the elements h_{ij} are all zero and we have a homomorphism to $\mathfrak{g}$ from a generalized Kac–Moody algebra, for which character and denominator formulas have been established. By construction of the monster Lie algebra in [3], which is also discussed later in this paper, the radical of the invariant form is zero, so that the following corollary will apply to this algebra.

Corollary 4.1. *Let* $\mathfrak{g}$ *be a Lie algebra satisfying the conditions in Theorem* 4.1. *If the radical of the form on* $\mathfrak{g}$ *is zero then there is a generalized Kac–Moody algebra* $\mathfrak{l}$ *such that* $\mathfrak{l}/\mathfrak{c} = \mathfrak{g}$, *where* $\mathfrak{c}$ *is the center of* $\mathfrak{l}$.

5. Free subalgebras of GKM algebras

5.1. Free Lie algebras

Denote by $L(X)$ the free Lie algebra on a set X, if W is a vector space with basis X then we also use the notation $L(W) = L(X)$. We will assume that X is finite or countably infinite. Let $v = (v_1, v_2, \ldots)$ be an m-tuple where $m \in \mathbb{Z}_+ \cup \infty$, and $v_i \in \mathbb{N}$ satisfy $v_i = 0$ for i sufficiently large. We will use the notation $|v| = \sum_{i=1}^{m} v_i$. If T_i are indeterminates, let

$$T^v = \prod_{i \in \mathbb{Z}_+} T_i^{v_i} \in \mathbb{R}[T_1, \ldots, T_m] \quad \text{or} \quad \mathbb{R}[T_1, T_2, \ldots].$$

We will consider gradings of the Lie algebra $L(X)$ of the following type: Let $\Delta = \mathbb{Z}^m$ and assign to each element $x \in X$ a degree $\alpha \in \Delta$ i.e., specify a map $\phi : X \to \Delta$. Let n_α be the number of elements of X of degree α and assume this is finite for all $\alpha \in \Delta$. This defines a grading of $L(X)$ which we denote $L(X) = \coprod_{(\alpha \in \Delta)} L^\alpha(X)$. Let $d(\alpha) = \dim L^\alpha(X)$. We assume that $d(\alpha)$ is finite.

An example of this type of grading is the *multigradation* from Bourbaki [5], which is defined as follows: Enumerate the set X so that $X = \{x_i\}_{i \in I}$ where $I = \mathbb{Z}_+$ or $\{1, \ldots, m\}$. Denote by $\Delta_{|X|}$ the group of $|X|-$tuples. Let $\sigma : X \to \Delta_{|X|}$ be defined by $\sigma(x_i) = \varepsilon_i = (0, \ldots, 0, 1, 0, \ldots)$ where the 1 appears in position i. The map σ defines the multigradation of $L(X)$.

Given such a grading determined by the map ϕ and the group $\Delta = \mathbb{Z}^m$, if $\Delta' = \mathbb{Z}^n$ and if $\psi : \Delta \to \Delta'$ is a homomorphism, then $\psi \circ \phi : X \to \Delta'$ also defines a grading of the Lie algebra $L(X)$. It is clear that if $\alpha \in \Delta'$ then $d(\alpha) = \sum_{\beta \in \Delta, \psi(\beta) = \alpha} d(\beta)$ as $L^\alpha(X) = \coprod_{\beta \in \Delta, \psi(\beta) = \alpha} L^\beta(X)$. Notice that $d(\alpha)$ is zero unless α has all nonnegative or all nonpositive entries, if the degree of each element of X has this property. The map ψ induces a homomorphism from $\mathbb{R}[T_1, \ldots, T_m]$ to $\mathbb{R}[T_1, \ldots T_n]$, where if $\alpha \in \Delta$ then $T^\alpha \mapsto T^{\psi(\alpha)}$.

Proposition 5.1. *Let* $L(X) = \coprod_{(\alpha \in \Delta)} L^\alpha(X)$ *be a grading of the above type. Then*

$$1 - \sum_{\alpha \in \Delta} n_\alpha T^\alpha = \prod_{\alpha \in \Delta \setminus \{0\}} (1 - T^\alpha)^{d(\alpha)}. \tag{11}$$

Proof. If we consider the multigradation of $L(X)$ by $\Delta_{|X|}$, so $L(X) = \coprod_{\beta\in\Delta_{|X|}} L^{\beta}(X)$, the following formula is proved in [5] (for finite X, but this immediately implies the result for an arbitrary X):

$$1 - \sum_{i\in I} T_i = \prod_{\beta\in\Delta_{|X|}\setminus\{0\}} (1 - T^{\beta})^{d(\beta)}. \tag{12}$$

This implies a formula for the more general type of grading given by a map $\phi : X \to \Delta$, as above. Any such map satisfies $\phi = \phi' \circ \sigma$ where $\phi' : \Delta_{|X|} \to \Delta$ is given by

$$\varepsilon_i \mapsto \phi(x_i).$$

Applying the homomorphism ϕ' to the identity (12) gives the proposition. □

Some of our results will follow from the elimination theorem in [5], which is restated here for the convenience of the reader.

Lemma 5.1 (Elimination theorem). *Let X be a set, S a subset of X and T the set of sequences $(s_1,\ldots,s_n,x)$ with $n \geq 0, s_1,\ldots,s_n$ in S and x in $X\setminus S$.*

(a) *The Lie algebra $L(X)$ is the direct sum as a vector space of the subalgebra $L(S)$ of $L(X)$ and the ideal $\mathfrak{a}$ of $L(X)$ generated by $X\setminus S$.*

(b) *There exists a Lie algebra isomorphism ϕ of $L(T)$ onto $\mathfrak{a}$ which maps $(s_1,\ldots,s_n,x)$ to $(\operatorname{ad} s_1 \cdots \operatorname{ad} s_n)(x)$.*

As Bourbaki [5] does for two particular gradings, we obtain formulas for computing the dimension of the homogeneous subspaces of a free Lie algebra $L(X)$ graded as above by a group Δ. The formulas derived here relate the dimension of the piece of degree α with the number of generators in that degree (which is assumed finite). If $\beta\in\Delta$ can be partitioned $\beta = \sum a_\alpha\alpha$, where $a_\alpha\in\mathbb{N}$ and $\alpha\in\Delta$, then define the partitions $P(\beta,j) = \{a = (a_\alpha)_{\alpha\in\Delta} | \beta = \sum a_\alpha\alpha, |a| = j\}$ and $P(\beta) = \bigcup_j P(\beta,j)$. Taking log on both sides of formula (11) leads to the equations

$$-\log\left(1 - \sum_{\alpha\in\Delta} n_\alpha T^\alpha\right) = \sum_{j\geq 1} \frac{1}{j}\left(\sum_{\alpha\in\Delta} n_\alpha T^\alpha\right)^j = \sum_{j\geq 1}\frac{1}{j}\sum_{\beta\in\Delta}\sum_{a\in P(\beta,j)} \frac{|a|!}{a!}\prod_{\alpha\in\Delta} n_\alpha^{a_\alpha} T^\beta$$

and

$$-\sum_{\alpha\in\Delta} d(\alpha)\log(1 - T^\alpha) = \sum_{\alpha\in\Delta,k\geq 1}\frac{1}{k}d(\alpha)T^{k\alpha} = \sum_{\beta}\sum_{k|\beta}\frac{1}{k}d(\beta/k)T^\beta.$$

Thus, if $\gamma|\beta$ means $k\gamma = \beta$ then

$$\sum_{\gamma|\beta}\frac{\gamma}{\beta}d(\gamma) = \sum_{a\in P(\beta)}\frac{(|a|-1)!}{a!}n^a$$

where n^a denotes the product of the $n_\alpha^{a_\alpha}$. Applying the Möbius inversion formula gives:

Proposition 5.2. *Let Δ be a grading of the free Lie algebra $L(X)$ as in Proposition 5.1. If $d(\beta)$ is the dimension of $L^{\beta}(X)$ for $\beta \in \Delta$ then*

$$d(\beta) = \sum_{\gamma|\beta} \left(\frac{\gamma}{\beta}\right) \mu\left(\frac{\beta}{\gamma}\right) \sum_{a \in P(\gamma)} \frac{(|a|-1)!}{a!} n^a. \tag{13}$$

5.2. Applications to generalized Kac–Moody algebras

In this section we will show that certain generalized Kac–Moody algebras contain large subalgebras which are isomorphic to free Lie algebras. We will apply the results of the preceding section to these examples.

We will begin with an easy example. Let A be a matrix satisfying conditions **(C1)**–**(C3)** which has no $a_{ii} > 0$, and all $a_{ij} < 0$ for $i \neq j$ (this means that imaginary simple roots are not mutually orthogonal). In this case, the generalized Kac–Moody algebra $\mathfrak{g}(A)$ is equal to $\mathfrak{g}_0(A)$. By Proposition 2.1, $\mathfrak{g} = \mathfrak{n}^+ \oplus \mathfrak{h} \oplus \mathfrak{n}^-$ where $\mathfrak{n}^{\pm}$ are the free Lie algebras on the sets $\{e_i\}_{i \in I}$ and $\{f_i\}_{i \in I}$, respectively.

Remark. Formula (11) can be applied to the root grading of $\mathfrak{g}(A)$ to obtain the denominator identity for $\mathfrak{g}(A)$. The root multiplicities are given by (13).

If the imaginary simple roots of a generalized Kac–Moody algebra are not mutually orthogonal then $\mathfrak{n}^{\pm}$ are not in general free, but we will show they contain ideals which are isomorphic to free Lie algebras. First we will set up some notation. Let $J \subset I$ be the set $\{i \in I \mid \alpha_i \in \Delta_R\} = \{i \in I \mid a_{ii} > 0\}$. Note that the matrix $(a_{ij})_{i,j \in J}$ is a generalized Cartan matrix, let $\mathfrak{g}_J$ be the Kac–Moody algebra associated to this matrix. Then $\mathfrak{g}_J = \mathfrak{n}_J^+ \oplus \mathfrak{h}_J \oplus \mathfrak{n}_J^-$, and $\mathfrak{g}_J$ is isomorphic to the subalgebra of $\mathfrak{g}(A)$ generated by $\{e_i, f_i\}$ with $i \in J$.

Theorem 5.1. *Let A be a matrix satisfying conditions* **(C1)**–**(C3)**. *Let J and $\mathfrak{g}_J$ be as above. Assume that if $i, j \in I \backslash J$ and $i \neq j$ then $a_{ij} < 0$. Then*

$$\mathfrak{g}(A) = \mathfrak{u}^+ \oplus (\mathfrak{g}_J + \mathfrak{h}) \oplus \mathfrak{u}^-,$$

where $\mathfrak{u}^- = L(\coprod_{j \in I \backslash J} \mathcal{U}(\mathfrak{n}_J^-) \cdot f_j)$ and $\mathfrak{u}^+ = L(\coprod_{j \in I \backslash J} \mathcal{U}(\mathfrak{n}_J^+) \cdot e_j)$. The $\mathcal{U}(\mathfrak{n}_J^-) \cdot f_j$ for $j \in I \backslash J$ are integrable highest weight $\mathfrak{g}_J$-modules, and the $\mathcal{U}(\mathfrak{n}_J^+) \cdot e_j$ are integrable lowest weight $\mathfrak{g}_J$-modules.

Note that the conditions on the a_{ij} given in the theorem are equivalent to the statement that the Lie algebra has no mutually orthogonal imaginary simple roots.

Proof. We will consider $\mathfrak{n}^+$; the case of $\mathfrak{n}^-$ is shown by a similar argument or by applying the automorphism η.

By the construction in Section 2, we have

$$\mathfrak{n}^+ = L(\{e_i\}_{i \in I})/\mathfrak{k}_0^+,$$

where $\mathfrak{k}_0^+$ is generated as an ideal of $L(\{e_i\}_{i\in I})$ by the elements

$$\{(\operatorname{ad} e_i)^{1-2a_{ij}/a_{ii}} e_j \mid i,j \in J, i \neq j\}$$

and

$$\{(\operatorname{ad} e_i)^{1-2a_{ij}/a_{ii}} e_j \mid i \in J, j \in I\backslash J\}.$$

This is because there are no elements of type (4). Apply the elimination theorem to the free Lie algebra $L(\{e_i\}_{i\in I})$ with $S = J$. Thus,

$$L(\{e_i\}_{i\in I}) = L(\{e_i\}_{i\in J}) \leftthreetimes \mathfrak{a},$$

where the ideal $\mathfrak{a}$ is isomorphic to the free Lie algebra on the set $X = \{\operatorname{ad} e_{i_1} \operatorname{ad} e_{i_2} \cdots \operatorname{ad} e_{i_k} e_j \mid j \in I\backslash J \text{ and } i_m \in J\}$. Let W denote the vector space with basis X, so that $\mathfrak{a} \cong L(W)$. Observe that as an $\mathfrak{h}^e$-module $W \cong \coprod_{j\in I\backslash J} \mathcal{U}(\mathfrak{l}) e_j$, where $\mathfrak{l}$ denotes the free Lie algebra $L(\{e_i\}_{i\in J})$.

For each fixed $j \in I\backslash J$ consider the submodule

$$R_j = \coprod_{i\in J} \mathcal{U}(\mathfrak{l})(\operatorname{ad} e_i)^{1-2a_{ij}/a_{ii}} e_j \subset \mathcal{U}(\mathfrak{l}) e_j.$$

Thus, identifying quotient spaces with subspaces of W,

$$\begin{aligned} W &= \coprod_{j\in I\backslash J} (\mathcal{U}(\mathfrak{l}) e_j / R_j \oplus R_j) \\ &= \coprod_{j\in I\backslash J} \mathcal{U}(\mathfrak{l}) e_j / R_j \oplus \coprod_{j\in I\backslash J} R_j. \end{aligned}$$

Now apply the elimination theorem to the Lie algebra $L(X) = L(W)$, choosing a basis of W of the form $S_1 \cup S_2$ where S_1 is a basis of the vector space $\coprod_{j\in I\backslash J} \mathcal{U}(\mathfrak{l}) e_j / R_j$ and S_2 is a basis of $\coprod_{j\in I\backslash J} R_j$. Obtaining

$$L(W) = L\left(\coprod_{j\in I\backslash J} \mathcal{U}(\mathfrak{l}) e_j / R_j \right) \leftthreetimes \mathfrak{b},$$

where $\mathfrak{b}$ is the ideal of $L(W)$ that is generated by S_2, i.e., by $\coprod_{j\in I\backslash J} R_j$. So

$$L\left(\{e_i\}_{i\in I}) = L(\{e_i\}_{i\in J}) \leftthreetimes L \coprod_{j\in I\backslash J} \mathcal{U}(\mathfrak{l}) e_j / R_j \right) \leftthreetimes \mathfrak{b}.$$

Let $\mathfrak{k}_J^+$ be the ideal of $\mathfrak{l}$ generated by

$$\{(\operatorname{ad} e_i)^{1-2a_{ij}/a_{ii}} e_j \mid i,j \in J, i \neq j\},$$

then $\mathcal{U}(\mathfrak{n}_J^+) = \mathcal{U}(\mathfrak{l}/\mathfrak{k}_J^+) = \mathcal{U}(\mathfrak{l})/\mathcal{K}$, where $\mathcal{K}$ denotes the ideal of $\mathcal{U}(\mathfrak{l})$ generated by $\mathfrak{k}_J^+ \subset \mathcal{U}(\mathfrak{l})$. Thus, we can decompose the vector space

$$\mathcal{U}(\mathfrak{l}) e_j / R_j = \mathcal{U}(\mathfrak{n}_J^+) e_j \Big/ \coprod_{i\in J} \mathcal{U}(\mathfrak{n}_J^+)(\operatorname{ad} e_i)^{1-2a_{ij}/a_{ii}} e_j \oplus \mathcal{K} e_j.$$

Applying the elimination theorem once again, using the above decomposition we obtain

$$L\left(\coprod_{j\in I\setminus J}\mathscr{U}(\mathfrak{l})e_j/R_j\right)=L\left(\coprod_{j\in I\setminus J}\left(\mathscr{U}(\mathfrak{n}_J^+)e_j/\coprod_{i\in J}\mathscr{U}(\mathfrak{n}_J^+)(\operatorname{ad}e_i)^{1-2a_{ij}/a_{ii}}e_j\right)\right)\rtimes\mathfrak{c},$$

where $\mathfrak{c}$ is the ideal in $L(\coprod_{j\in I\setminus J}\mathscr{U}(\mathfrak{l})e_j/R_j)$ generated by the sum of the $\mathscr{K}e_j$. Each $\mathfrak{h}^e$-module $\mathscr{U}(\mathfrak{n}_J^+)e_j/\coprod_{i\in J}\mathscr{U}(\mathfrak{n}_J^+)(\operatorname{ad}e_i)^{1-2a_{ij}/a_{ii}}e_j$ is an integrable lowest weight module for the Lie algebra $\mathfrak{g}_J$, denoted by $\mathscr{U}(\mathfrak{n}_J^+)\cdot e_j$, with lowest weight α_j. Thus, we have a decomposition into semidirect products:

$$L(\{e_i\}_{i\in I})=L(\{e_i\}_{i\in J})\rtimes\left[\left(L\left(\coprod_{i\in I\setminus J}\mathscr{U}(\mathfrak{n}_J^+)\cdot e_j\right)\rtimes\mathfrak{c}\right)\rtimes\mathfrak{b}\right].$$

It is clear that, as ideals of $L(\{e_i\}_{i\in I})$, $\mathfrak{b},\mathfrak{c}\subset\mathfrak{k}_0^+$, and $\mathfrak{k}_J^+$ is $\mathfrak{k}_0^+\cap L(\{e_i\}_{i\in J})$. Therefore, since all elements of $\mathfrak{k}_0^+$ are zero in $L(\coprod_{i\in I\setminus J}\mathscr{U}(\mathfrak{n}_J^+)\cdot e_j)$,

$$L(\{e_i\}_{i\in I})/\mathfrak{k}_0^+=L(\{e_i\}_{i\in J})/\mathfrak{k}_J^+\rtimes L\left(\coprod_{j\in I\setminus J}\mathscr{U}(\mathfrak{n}_J^+)\cdot e_j\right).$$

By the definition of $\mathfrak{g}_J$, $\mathfrak{n}_J^+=L(\{e_i\}_{i\in J})/\mathfrak{k}_J^+$. □

Corollary 5.1. *Let A be a matrix satisfying conditions* **(C1)**–**(C3)**. *Assume that the matrix A has only one positive diagonal entry, $a_{ii}>0$, and if $a_{mj}=0$ then $m=i$, or $j=i$ or $m=j$. Let $S=\{(\operatorname{ad}e_i)^l e_j\}_{0\le l\le -2a_{ij}/a_{ii}}$. The subalgebra $\mathfrak{n}^+\subset\mathfrak{g}(A)$ is the semidirect product of a one-dimensional Lie algebra and a free Lie algebra, $\mathfrak{n}^+=\mathbb{R}e_i\oplus L(S)$. Similarly, $\mathfrak{n}^-=\mathbb{R}f_i\oplus L(\eta(S))$. Thus,*

$$\mathfrak{g}(A)=L(S)\oplus(\mathfrak{sl}_2+\mathfrak{h})\oplus L(\eta(S)).$$

The root grading is a grading of the type considered in Proposition 5.1 because we have the correspondence $\alpha_i\mapsto(0,\dots,1,\dots,0)$, where 1 appears in the ith place. This is the grading (6). The denominator formula given in the next result is the same as (7), after the change of variables $e^\alpha\mapsto T^\alpha$.

Corollary 5.2. *Let A be as in Theorem* 5.1, *n_α and T^α as in Proposition* 5.1. *Let $\Delta_+^J=\Delta\cap\sum_{i\in J}\mathbb{Z}_+\alpha_i$. The Lie algebra $\mathfrak{g}(A)$ has denominator formula given by* (11) *and the denominator formula for $\mathfrak{g}_J$:*

$$\begin{aligned}&\prod_{\varphi\in\Delta_+}(1-T^\varphi)^{\dim\mathfrak{g}^\varphi}\\&=\prod_{\varphi\in\Delta_+^J}(1-T^\varphi)^{\dim\mathfrak{g}_J^\varphi}\cdot\prod_{\varphi\in\Delta_+\setminus\Delta_+^J}(1-T^\varphi)^{\dim L^\varphi\left(\coprod_{j\in I\setminus J}\mathscr{U}(\mathfrak{n}_J^+)\cdot e_j\right)}\\&=\left(\sum_{w\in W}(-1)^{l(w)}T^{w\rho-\rho}\right)\left(1-\sum_{\varphi\in\Delta^+\setminus\Delta_+^J}n_\varphi T^\varphi\right).\end{aligned}$$

Obtaining the denominator formula in this way provides an alternative proof that the radical of the Lie algebra $\mathfrak{g}$ associated to the matrix A is zero. This is because Corollary 5.2 uses only the description of $\mathfrak{g}$ in terms of generators and relations, while the previous proof of the denominator formula (see [17] or [18]) is valid after the radical of the Lie algebra has been factored out. Thus, we have shown, for the particular type of matrix in Theorem 5.1:

Corollary 5.3. *Let A be as in Theorem 5.1. The generalized Kac–Moody algebra $\mathfrak{g}(A)$ has zero radical.*

Remark. A proof of the fact that any generalized Kac–Moody algebra has zero radical can be found in [16] or [17]. In both cases the argument of [12] or [18] is extended to include generalized Kac–Moody algebras by making use of a lemma appearing in [2].

Remark. If we apply (13) to $L(\coprod_{j\in I\setminus J}\mathscr{U}(\mathfrak{n}_j^+)\cdot e_j)$ we obtain Kang's [19] multiplicity formulas for the special case of generalized Kac–Moody algebras with no mutually orthogonal imaginary simple roots.

5.3. A Lie algebra related to the modular function j

We will apply our results to an important example of a generalized Kac–Moody algebra $\mathfrak{g}(M)$, defined in Section 6.2 below. It will be shown that, if $\mathfrak{c}$ denotes the center of the Lie algebra $\mathfrak{g}(M)$, then $\mathfrak{g}(M)/\mathfrak{c}$ is the Monster Lie algebra.

Recall that the modular function j has the expansion $j(q)=\sum_{i\in\mathbb{Z}}c(i)q^i$, where $c(i)=0$ if $i<-1$, $c(-1)=1$, $c(0)=744$, $c(1)=196\,884$. Let $J(v)=\sum_{i\geq -1}c(i)v^i$ be the formal Laurent series associated to $j(q)-744$.

Let M be the symmetric matrix of blocks indexed by $\{-1,1,2,\ldots\}$, where the block in position (i,j) has entries $-(i+j)$ and size $c(i)\times c(j)$. Thus

$$M=\left(\begin{array}{c|ccc|ccc|c}
2 & 0 & \cdots & 0 & -1 & \cdots & -1 & \cdots \\
\hline
0 & -2 & \cdots & -2 & -3 & \cdots & -3 & \\
\vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \cdots \\
0 & -2 & \cdots & -2 & -3 & \cdots & -3 & \\
\hline
-1 & -3 & \cdots & -3 & -4 & \cdots & -4 & \\
\vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \cdots \\
-1 & -3 & \cdots & -3 & -4 & \cdots & -4 & \\
\hline
\vdots & & \vdots & & & \vdots & &
\end{array}\right).$$

Definition 3. Let $\mathfrak{g}(M)$ be the generalized Kac–Moody algebra associated to the matrix M given above.

We have the standard decomposition

$$\mathfrak{g}(M) = \mathfrak{n}^+ \oplus \mathfrak{h} \oplus \mathfrak{n}^-. \tag{14}$$

The generators of $\mathfrak{g}(M)$ will be written e_{jk}, f_{jk}, h_{jk}, indexed by integers j,k where $j \in \{-1\} \cup \mathbb{Z}_+$ and $1 \leq k \leq c(j)$. Since there is only one e, f or h with $j = -1$, we will write these elements as e_{-1}, f_{-1}, h_{-1}. From the construction of $\mathfrak{g}(M)$ we see that the simple roots in $(\mathfrak{h}^e)^*$ are $\alpha_{-1}, \alpha_{11}, \alpha_{12}, \ldots, \alpha_{1c(1)}, \alpha_{21}, \ldots, \alpha_{2c(2)}$, etc. Note that for fixed i the functionals α_{ij} and α_{ik} agree on all of $\mathfrak{h}$ for $1 \leq j,k \leq c(i)$. The α_{ik} for $i > -1$ are simple imaginary roots, and the root α_{-1} is the one real simple root.

Remark. We explain the relationship between our definition of simple root and that appearing in [4]. If the restrictions of the simple roots to $\mathfrak{h}$ are denoted $\alpha_{-1}, \alpha_1, \alpha_2, \ldots$, these elements of $\mathfrak{h}^*$ correspond to the notion of "simple imaginary roots of multiplicity greater than one" in [4]. The "simple root" α_i has "multiplicity" $c(i)$ in Borcherds' terminology. Fortunately, in this case, nonsimple roots $\alpha \in (\mathfrak{h}^e)^*$ do not restrict to any α_i, also "multiplicities" do not become infinite, "roots" remain either positive or negative, etc. The functionals α_i are linearly dependent, in fact they span a two-dimensional space. The root lattice is described in [4] as the lattice $\mathbb{Z} \oplus \mathbb{Z}$ with the inner product given by

$$\begin{pmatrix} 0 & -1 \\ -1 & 0 \end{pmatrix}.$$

The "simple roots" of this Lie algebra are denoted $(1,n)$ where $n = -1$ or $n \in \mathbb{Z}_+$. However, in most cases (including the case of $\widehat{\mathfrak{sl}_2}$) serious problems arise when we do not work in a sufficiently large Cartan subalgebra. (We do not wish to write down a "denominator identity" where some of the terms are ∞.) Here, we always work in $(\mathfrak{h}^e)^*$, taking specializations when they are illuminating, as in the case of the denominator identity for $\mathfrak{g}(M)$ given below.

Corollary 5.1 applied to the Lie algebra $\mathfrak{g}(M)$ gives the following.

Theorem 5.2. *The subalgebra $\mathfrak{n}^+ \subset \mathfrak{g}(M)$ is the semidirect product of a one-dimensional Lie algebra and a free Lie algebra, so $\mathfrak{n}^+ = \mathbb{R}e_{-1} \oplus L(S)$. Similarly, $\mathfrak{n}^- = \mathbb{R}f_{-1} \oplus L(S')$. Hence,*

$$\mathfrak{g}(M) = L(S) \oplus (\mathfrak{sl}_2 + \mathfrak{h}) \oplus L(S').$$

Here $S = \bigcup_{j \in \mathbb{N}} \{(\operatorname{ad} e_{-1})^l e_{jk} \mid 0 \leq l < j, 1 \leq k \leq c(j)\}$, and $S' = \eta(S)$.

Now that we have established that $\mathfrak{n}^+$ is the direct sum of a one dimensional space and an ideal isomorphic to a free Lie algebra we shall obtain the denominator formula for the Lie algebra $\mathfrak{g}(M)$.

Corollary 5.4. *The denominator formula for the Lie algebra* $\mathfrak{g}(M)$ *is*

$$\prod_{\varphi\in\Delta_+}(1-T^{\varphi})^{\dim\mathfrak{g}^{\varphi}} = (1-T^{\alpha_{-1}}) \prod_{\varphi\in\Delta_+\setminus\{\alpha_{-1}\}}(1-T^{\varphi})^{\dim L^{\varphi}(S)}$$

$$= (1-T^{\alpha_{-1}})\left(1-\sum_{\substack{j\in\mathbb{Z}_+\\ 1\le k\le a(j),0\le l<j}} T^{l\alpha_{-1}+\alpha_{jk}}\right), \tag{15}$$

which has specialization

$$u(J(u)-J(v)) = \prod_{\substack{i\in\mathbb{Z}_+\\ j\in\mathbb{Z}_+\cup\{-1\}}}(1-u^iv^j)^{c(ij)} \tag{16}$$

under the map $\phi : \Delta \to \mathbb{Z}\times\mathbb{Z}$ *determined by* $\alpha_{ik}\mapsto(1,i)$, *where we write* $T^{\alpha_{ik}}\mapsto uv^i$.

Proof of Corollary 5.4. For the denominator identity simply apply Corollary 5.2 to the root grading. The $\mathbb{Z}\times\mathbb{Z}$-grading of $L(S)$ given above is such that a generator $(\operatorname{ad}e_{-1})^l e_{jk}$ has degree $l(1,-1)+(1,j)=(l+1,j-l)$ with $l<j$. The number of generators of degree (i,j) is $c(i+j-1)$. Applying Eq. 11 (which is the same as specializing the second product of the denominator identity via $T^{\alpha_i}\mapsto uv^i$) gives the formula

$$1-\sum_{(i,j)\in\mathbb{N}^2\setminus\{0\}} c(i+j-1)u^iv^j = \prod_{(i,j)\in\mathbb{N}^2\setminus\{0\}}(1-u^iv^j)^{\dim L^{(i,j)}}.$$

To obtain the specialization of the denominator formula of $\mathfrak{g}$ we must include the degree $(1,-1)$ subspace $\mathfrak{g}^{(1,-1)}=\mathbb{R}e_{-1}$, which is one-dimensional:

$$\prod_{(i,j)}(1-u^iv^j)^{\dim\mathfrak{g}^{(i,j)}} = \prod_{(i,j)\in\mathbb{N}^2-\{0\}}(1-u^iv^j)^{\dim L^{(i,j)}}(1-u/v)$$

$$= \left(1-\sum_{(i,j)\in\mathbb{N}^2-\{0\}} c(i+j-1)u^iv^j\right)(1-u/v)$$

$$= 1-\sum c(i+j-1)u^iv^j - u/v + \sum c(i+j-1)u^{i+1}v^{j-1}$$

$$= u(J(u)-J(v)).$$

There is a product formula for the modular function j (see [4]) which can be written:

$$p(j(p)-j(q)) = \prod_{\substack{i=1,2,\dots\\ j=-1,1,\dots}}(1-p^iq^j)^{c(ij)}, \tag{17}$$

which converges on an open set in $\mathbb{C}$, and so implies the corresponding identity for formal power series. Now we conclude that $\dim\mathfrak{g}^{(i,j)}=c(ij)$. □

Note that here we must know the number theory identity of [4], to determine the dimension of the root spaces of $\mathfrak{g}$.

The identity (16) is the specialization $e^{\alpha_i} \mapsto uv^i$ of the denominator identity as it appears when we apply Eq. (7) to $\mathfrak{g}(M)$.

Remark. The matrix M can be replaced by any symmetric matrix with the same first row (and column) as M with all remaining blocks having entries strictly less than zero, as long as the minor obtained by removing the first row and column has the same rank as the corresponding minor of M.

Remark. Now we apply Eq. (13) to the $\mathbb{N} \times \mathbb{N}$-grading, and the Lie algebra $L(S)$, where $S = \{(\operatorname{ad} e_{-1})^l e_{jk} \mid 0 \leq l < j, 1 \leq k \leq c(j)\}$ and there are $c(i+j-1)$ generators of degree (i,j). Since we already know the dimension of $L^{(i,j)}$ is $c(ij)$, we recover (see [19]) the following relations between the coefficients of j:

$$c(ij) = \sum_{\substack{k \in \mathbb{Z}_+ \\ k(m,n)=(i,j)}} \frac{1}{k}\mu(k) \sum_{a \in P(m,n)} \frac{(\sum a_{rs} - 1)!}{\prod a_{rs}!} \prod c(r+s-1)^{a_{rs}}. \tag{18}$$

6. The Monster Lie algebra

6.1. Vertex operator algebras and vertex algebras

For a detailed discussion of vertex operator algebras and vertex algebras the reader should consult [8, 9, 11] and the announcement [1]. Results stated here without proof can either be found in [8, 9, 11] or follow without too much difficulty from the results appearing there.

Definition 4. A *vertex operator algebra*, $(V, Y, \mathbf{1}, \omega)$, consists of a vector space V, distinguished vectors called the *vacuum vector* $\mathbf{1}$ and the *conformal vector* ω, and a linear map $Y(\cdot,z) : V \to (\operatorname{End} V)[[z, z^{-1}]]$ which is a generating function for operators v_n, i.e., for $v \in V$, $Y(v,z) = \sum_{n \in \mathbb{Z}} v_n z^{-n-1}$, satisfying the following conditions:

(V1) $V = \coprod_{n \in \mathbb{Z}} V_{(n)}$; for $v \in V_{(n)}$, $n = \operatorname{wt}(v)$,

(V2) $\dim V_{(n)} < \infty$ for $n \in \mathbb{Z}$,

(V3) $V_{(n)} = 0$ for n sufficiently small,

(V4) If $u, v \in V$ then $u_n v = 0$ for n sufficiently large,

(V5) $Y(\mathbf{1},z) = 1$,

(V6) $Y(v,z)\mathbf{1} \in V[[z]]$ and $\lim_{z \to 0} Y(v,z)\mathbf{1} = v$, i.e., the *creation property* holds,

(V7) the following *Jacobi identity* holds:

$$z_0^{-1}\delta\left(\frac{z_1 - z_2}{z_0}\right) Y(u,z_1)Y(v,z_2) - z_0^{-1}\delta\left(\frac{z_2 - z_1}{-z_0}\right) Y(v,z_2)Y(u,z_1)$$

$$= z_2^{-1}\delta\left(\frac{z_1 - z_0}{z_2}\right) Y(Y(u,z_0)v,z_2). \tag{19}$$

The following conditions relating to the vector ω also hold:

(V8) The operators ω_n generate a Virasoro algebra, i.e., if we let $L(n) = \omega_{n+1}$ for $n \in \mathbb{Z}$ then

$$[L(m), L(n)] = (m-n)L(m+n) + (1/12)(m^3 - m)\delta_{m+n,0}(\operatorname{rank} V), \tag{20}$$

(V9) If $v \in V_{(n)}$ then $L(0)v = (\operatorname{wt} v)v = nv$,

(V10) $(\mathrm{d}/\mathrm{d}z)Y(v,z) = Y(L(-1)v,z)$.

Definition 5. A *vertex algebra* $(V, Y, 1, \omega)$ is a vector space V with all of the above properties except for **(V2)** and **(V3)**.

Remark. This definition is a variant, with ω, of Borcherds' original definition of vertex algebra in [1].

An important class of examples of vertex algebras (and vertex operator algebras) are those associated with lattices. For the sake of the reader who may be unfamiliar with the notation we will briefly review this construction in the case of an even lattice. For complete details (and more generality) the reader may consult [11] or [8]. Given an even lattice L one can construct a vertex algebra V_L with underlying vector space:

$$V_L = S(\hat{\mathfrak{h}}_{\mathbb{Z}}^{-}) \otimes \mathbb{R}\{L\}.$$

Here we take $\mathfrak{h} = L \otimes_{\mathbb{Z}} \mathbb{R}$, and $\hat{\mathfrak{h}}_{\mathbb{Z}}^{-}$ is the negative part of the Heisenberg algebra (with c central) defined by

$$\hat{\mathfrak{h}}_{\mathbb{Z}} = \coprod_{n \in \mathbb{Z}} \mathfrak{h} \otimes t^n \oplus \mathbb{R}c \subset \mathfrak{h} \otimes \mathbb{R}[[t]] \oplus \mathbb{R}c.$$

Therefore,

$$\hat{\mathfrak{h}}_{\mathbb{Z}}^{-} = \coprod_{n<0} \mathfrak{h} \otimes t^n.$$

The symmetric algebra on $\hat{\mathfrak{h}}_{\mathbb{Z}}^{-}$ is denoted $S(\hat{\mathfrak{h}}_{\mathbb{Z}}^{-})$. Given a central extension of L by a group of order 2, i.e.,

$$1 \to \langle \kappa | \kappa^2 = 1 \rangle \to \hat{L} \to L \to 1,$$

with commutator map given by $\kappa^{\langle \alpha, \beta \rangle}$, $\alpha, \beta \in L$, $\mathbb{R}$ is given the structure of a nontrivial $\langle \kappa \rangle$-module. Define $\mathbb{R}\{L\}$ to be the induced representation $\operatorname{Ind}_{\langle \kappa \rangle}^{\hat{L}} \mathbb{R}$.

If $a \in \hat{L}$ denote by $\iota(a)$ the element $a \otimes 1 \in \mathbb{R}\{L\}$. We will use the notation $\alpha(n) = \alpha \otimes t^n \in S(\hat{\mathfrak{h}}_{\mathbb{Z}}^{-})$. The vector space V_L is spanned by elements of the form

$$\alpha_1(-n_1)\alpha_2(-n_2)\ldots\alpha_k(-n_k)\iota(a),$$

where $n_i \in \mathbb{N}$. The space V_L, equipped with $Y(v,z)$ as defined in [11] satisfies properties **(V1)** and **(V4)**–**(V10)**, so is a vertex algebra with conformal vector ω. Features of $Y(v,z)$ to keep in mind from [11] are: $\alpha(-1)_n = \alpha(n)$ for all $n \in \mathbb{Z}$; the $\alpha(n)$ for $n < 0$ act by left multiplication on $u \in V_L$, and

$$\alpha(n)1(a) = \begin{cases} 0 & \text{if } n>0, \\ \langle \alpha, \bar{a} \rangle 1(a) & \text{if } n = 0. \end{cases}$$

Definition 7. A bilinear form $(\cdot,\cdot)$ on a vertex algebra V is *invariant* (in the sense of [9]) if it satisfies

$$(Y(v,z)w_1, w_2) = (w_1, Y(e^{zL(1)}(-z^{-2})^{L(0)}v, z^{-1})w_2). \tag{21}$$

(By definition $x^{L(0)}$ acts on a homogeneous element $v \in V$ as multiplication by $x^{wt(v)}$.) Such a form satisfies $(u,v) = 0$ unless $\mathrm{wt}(u) = \mathrm{wt}(v)$.

Lemma 6.1. *Let L be an even unimodular lattice. There is a nondegenerate symmetric invariant bilinear form $(\cdot,\cdot)$ on V_L.*

Proof. The vertex algebra V_L is a module for itself under the adjoint action. In fact, V_L is an irreducible module and any irreducible module of V_L is isomorphic to V_L [7].

In order to define the contragredient module note that V_L is graded by the lattice L as well as by weights, and that under this double grading $\dim V^r_{L_{(n)}} < \infty$ for $r \in L$, $n \in \mathbb{Z}$. Let $V_L' = \coprod_{n\in\mathbb{Z}, r\in L}(V^r_{L_{(n)}})^*$, the restricted dual of V_L. Denote by $\langle \cdot,\cdot \rangle$ the natural pairing between V_L and V_L'. Results of [9] pertaining to adjoint vertex operators and the contragredient module now apply to V_L'. In particular, the space V_L' can be given the structure of a V_L-module (V_L', Y') via

$$\langle Y'(v,z)w', w\rangle = \langle w', Y(e^{zL(1)}(-z^{-2})^{L(0)}v, z^{-1})w\rangle$$

for $v, w \in V_L$ $w' \in V_L'$. Since the adjoint module V_L is irreducible, the contragredient module V_L' is also irreducible. By the result of [7] quoted above, V_L is isomorphic to V_L' as a V_L-module, which is equivalent to V_L having a nondegenerate invariant bilinear form (see [9, Remark 5.3.3]). □

The "moonshine module" $V^\natural$ is an infinite-dimensional representation of the Monster simple group constructed and shown to be a vertex operator algebra in [11]. The graded dimension of $V^\natural$ is $J(q)$. There is a positive definite bilinear form $(\cdot,\cdot)$ on $V^\natural$. The vertex operator algebra $V^\natural$ satisfies all of the conditions of the no-ghost theorem (Theorem 6.2), with G taken to be the Monster simple group. This vertex operator algebra will be essential to the construction of the Monster Lie algebra.

Lemma 6.2. *The positive-definite form $(\cdot,\cdot)$ on $V^\natural$ defined in* [11] *is invariant.*

Proof. There is a unique up to constant multiple nondegenerate symmetric invariant bilinear form on $V^\natural$ [22]. Fix such a form $(\cdot,\cdot)_1$ by taking $(1,1)_1 = 1$.

Let $u \in V^{\natural}_{(2)}$, a homogeneous element of weight 2. By invariance

$$(u_n w_1, w_2)_1 = (w_1, u_{-n+2} w_2)_1$$

for $w_1, w_2 \in V^{\natural}$. We claim

$$(u_n w_1, w_2) = (w_1, u_{-n+2} w_2) \tag{22}$$

$w_1, w_2 \in V^{\natural}$. In order to prove Eq. (22) we recall the construction and properties of $V^{\natural}$ of [11]. Let $x_a^+ = \imath(a) + \imath(a^{-1})$ for $a \in \Lambda$ (the Leech lattice),

$$\mathfrak{k} = S^2(\mathfrak{h} \otimes t^{-1}) \oplus \sum_{a \in \hat{\Lambda}_4} \mathbb{R} x_a^+,$$

and let $\mathfrak{p}$ be the space of elements of V_λ^T (the "twisted space") of weight 2. Then $V^{\natural}_{(2)} = \mathfrak{k} \oplus \mathfrak{p}$. The action $Y(v,z)$ of elements $v \in \mathfrak{p}$ is determined by conjugating by certain elements of $\mathbb{M}$ the Monster simple group (see 12.3.8 and 12.3.9 of [11]). Conjugation by these elements map $v \in \mathfrak{p}$ to $\mathfrak{k}$. Since the form $(\cdot,\cdot)$ is invariant under $\mathbb{M}$, it is sufficient to check that Eq. (22) holds for elements of $\mathfrak{k}$. Therefore, it suffices to check Eq. (22) for two types of elements $x_a^+, a \in \hat{\Lambda}_4$ and $g(-1)^2, g(-1) \in \mathfrak{h} \otimes t^{-1}$. For $u = x_a^+$ Eq. (22) follows immediately from Proposition 12.5.1 of [11]. For $u = g(-1)^2$

$$\begin{aligned} Y(u,z) &= {}^{\circ}_{\circ} g(z)^2 {}^{\circ}_{\circ} \\ &= g(z)^- g(z) + g(z) g(z)^+. \end{aligned}$$

Using $(g(i)w_1, w_2) = (w_1, g(-i)w_2)$ one computes the adjoint of $g(z)^- g(z) + g(z)g(z)^+$ which is

$$g(z^{-1})g(z^{-1})^+ + g(z^{-1})^- g(z^{-1}) = Y(u, z^{-1}) z^{-4}.$$

Thus, Eq. (22) holds for all $u \in \mathfrak{k}$, and so for all $u \in V^{\natural}_{(2)}$.

Now recall that $V^{\natural}_{(2)} = \mathscr{B}$, the Griess algebra, and the notation $\hat{\mathscr{B}}$ for the commutative affinization of the algebra $\mathscr{B}$,

$$\hat{\mathscr{B}} = \mathscr{B} \otimes \mathbb{R}[t, t^{-1}] \oplus \mathbb{R}e,$$

where t is an indeterminate and $e \neq 0$ (with nonassociative product given in [11]). By Theorem 12.3.1 [11] $V^{\natural}$ is an irreducible graded $\hat{\mathscr{B}}$-module, under

$$\pi : \hat{\mathscr{B}} \to \operatorname{End} V^{\natural},$$
$$v \otimes t^n \mapsto x_v(n) \quad v \in \mathscr{B},$$
$$e \mapsto 1.$$

Schur's lemma then implies that any nondegenerate symmetric bilinear form satisfying Eq. (22) is unique up to multiplication by a constant. Thus, we can conclude that $(\cdot,\cdot)_1 = (\cdot,\cdot)$, since the length of the vacuum is one with respect to each form. □

Given V a vertex operator algebra, or a vertex algebra with ω and therefore an action of the Virasoro algebra, let

$$P_{(i)} = \{v \in V \mid L(0)v = iv, L(n)v = 0 \text{ if } n > 0\}.$$

Thus, $P_{(i)}$ consists of the lowest weight vectors for the Virasoro algebra of weight i. Then $P_{(1)}/L(-1)P_{(0)}$ is a Lie algebra with bracket given by $[u + L(-1)P_{(0)}, v+L(-1)P_{(0)}] = u_0v+L(-1)P_{(0)}$. If the vertex algebra V has an invariant bilinear form $(\cdot,\cdot)$ this induces a form $(\cdot,\cdot)_{Lie}$ on the Lie algebra $P_1/L(-1)P_{(0)}$, because $L_{-1}P_{(0)} \subset \mathrm{rad}(\cdot,\cdot)$. Invariance of the form on the vertex algebra implies for $u, v \in P_{(1)}$:

$$(u_0v, w) = -(v, u_0w). \tag{23}$$

So that the induced form is invariant on the Lie algebra $P_{(1)}/L(-1)P_{(0)}$.

Tensor products of vertex operator algebras are again vertex operator algebras (see [9]), and more generally, by [8], the tensor product of vertex algebras is also a vertex algebra. Given two vertex algebras $(V, Y, \mathbf{1}_V, \omega_V)$ and $(W, Y, \mathbf{1}_W, \omega_W)$ the vacuum of $V \otimes W$ is $\mathbf{1}_V \otimes \mathbf{1}_W$ and the conformal vector ω is given by $\omega_V \otimes \mathbf{1}_W + \mathbf{1}_V \otimes \omega_W$. If the vertex algebras V and W both have invariant forms then it is not difficult to show that the form on $V \otimes W$ given by the product of the forms on V and W is also invariant in the sense of Eq. (21).

6.2. *The Monster Lie algebra*

We will review the construction of the Monster Lie algebra given in [4]. Then we give a theorem regarding its structure as a quotient of $\mathfrak{g}(M)$. Let $\mathscr{L} = \mathbb{Z} \oplus \mathbb{Z}$ with bilinear form $\langle\cdot,\cdot\rangle$ given by the matrix

$$\begin{pmatrix} 0 & -1 \\ -1 & 0 \end{pmatrix}.$$

Remark. $\mathscr{L}$ is the rank two Lorentzian lattice, denoted in [4] as $II_{1,1}$.

Fix a symmetric invariant bilinear form $(\cdot,\cdot)$ on $V_{\mathscr{L}}$, normalized by taking $(\mathbf{1},\mathbf{1}) = -1$. The reason that we choose this normalization is that the resulting invariant bilinear form on the Monster Lie algebra will have the usual values with respect to the Chevalley generators, and so that the contravariant bilinear form defined below will be positive definite and not negative definite on nonzero weight spaces. Denote by $(\cdot,\cdot)$ the symmetric invariant bilinear form on $V^\natural \otimes V_{\mathscr{L}}$ given by the product of the invariant bilinear forms on $V^\natural$ and $V_{\mathscr{L}}$.

Definition 7. The *Monster Lie algebra* $\mathfrak{m}$ is defined by

$$\mathfrak{m} = P_1/\mathrm{rad}(\cdot,\cdot)_{Lie} = (P_1/L_{-1}P_0)/\mathrm{rad}(\cdot,\cdot)_{Lie}.$$

When no confusion will arise, we will use the same notation for the invariant form on the vertex algebra, and for the induced form on the Lie algebra.

Note that, by invariance $(e^r, e^s) = (-1)^{\langle r,r\rangle/2}\langle 1, (e^r)_{(-1+\langle r,r\rangle)}e^s\rangle = 0$ unless $r = -s \in \mathscr{L}$. Therefore, the induced form on $\mathfrak{m}$ satisfies the condition that $\mathfrak{m}_r$ be orthogonal to $\mathfrak{m}_s$ if $r \neq -s \in \mathscr{L}$.

Definition 8. Let θ be the involution of $V_{\mathscr{L}}$ given by $\theta\imath(a) = (-1)^{wt(a)}\imath(a^{-1})$ and $\theta(\alpha(n)) = -\alpha(n)$. This induces an involution θ on all of $V^\natural \otimes V_{\mathscr{L}}$ by letting $\theta(u \otimes ve^r) = u \otimes \theta(ve^r)$. Use the same notation for the involution induced by θ on $\mathfrak{m}$.

Note that $\theta : \mathfrak{m}_r \to \mathfrak{m}_{-r}$ if $r \neq 0$.

Let $(\cdot,\cdot)_0$ be the contravariant bilinear form on $V^\natural \otimes V_{\mathscr{L}}$ given by $(u,v)_0 = -(u, \theta(v))$, $u, v \in V^\natural \otimes V_{\mathscr{L}}$. We also denote by $(\cdot,\cdot)_0$ the contravariant bilinear form on $\mathfrak{m}$ given by $(u,v)_0 = -(u, \theta(v))_{Lie}$, $u, v \in \mathfrak{m}$.

Elements of $\mathfrak{m}$ can be written as $\sum u \otimes ve^r$, where $u \in V^\natural$ and $ve^r = v\imath(e^r) \in V_{\mathscr{L}}$. Here, a section of the map $\hat{\mathscr{L}} \to \mathscr{L}$ has been chosen so that $e^r \in \hat{\mathscr{L}}$ satisfies $\overline{e^r} = r \in \mathscr{L}$. There is a grading of $\mathfrak{m}$ by the lattice defined by $\deg(u \otimes ve^r) = r$.

Recall the definition of the Lie algebra $\mathfrak{g}(M)$ and the standard decomposition.

Theorem 6.1. *Let* $\mathfrak{c}$ *denote the center of the Lie algebra* $\mathfrak{g}(M)$. *Then*

$$\mathfrak{g}(M)/\mathfrak{c} = \mathfrak{m}.$$

There is a triangular decomposition $\mathfrak{m} = \mathfrak{m}^+ \oplus \mathfrak{h} \oplus \mathfrak{m}^-$, *where* $\mathfrak{h} \cong \mathbb{R} \oplus \mathbb{R}$. *The subalgebras* $\mathfrak{m}^\pm$ *are isomorphic to* $\mathfrak{n}^\pm \subset \mathfrak{g}(M)$.

Theorem 6.1 is proved after the statement of the no-ghost theorem. In [4] the no-ghost theorem from string theory is used to see that the Monster Lie algebra has homogeneous subspaces isomorphic to $V^\natural_{(1+mn)}$. A precise statement of the no-ghost theorem as it is used here is provided for the reader, and a proof is given in the appendix.

Theorem 6.2 (No-ghost theorem). *Let* V *be a vertex operator algebra with the following properties*:

(i) V *has a symmetric invariant nondegenerate bilinear form.*

(ii) *The central element of the Virasoro algebra acts as multiplication by* 24.

(iii) *The weight grading of* V *is an* $\mathbb{N}$*-grading of* V, *i.e.*, $V = \coprod_{n=0}^{\infty} V_{(n)}$, *and dim* $V_{(0)} = 1$.

(iv) V *is acted on by a group* G *preserving the above structure*; *in particular the form on* V *is* G*-invariant.*

Let $\mathscr{P}_{(1)} = \{u \in V \otimes V_{\mathscr{L}} \mid L_0 u = u, L_i u = 0, i > 0\}$. *The group* G *acts on* $V \otimes V_{\mathscr{L}}$ *via the trivial action on* $V_{\mathscr{L}}$. *Let* $\mathscr{P}^r_{(1)}$ *denote the subspace of* $\mathscr{P}_{(1)}$ *of degree* $r \in \mathscr{L}$. *Then the quotient of* $\mathscr{P}^r_{(1)}$ *by the nullspace of its bilinear form is isomorphic as a* G*-module with* G*-invariant bilinear form to* $V_{(1-\langle r,r\rangle/2)}$ *if* $r \neq 0$ *and to* $V_{(1)} \oplus \mathbb{R}^2$ *if* $r = 0$.

Proof of Theorem 6.1. The no-ghost theorem applied to $V^\natural$ immediately gives $\mathfrak{m}_{(m,n)} \cong V^\natural_{(mn+1)}$ if $(m,n) \neq (0,0)$. Thus, the elements of $\mathfrak{m}_r$ where $r \neq 0$ are spanned by

elements of the form $u \otimes e^r$ where $r \in \mathscr{L}$, and $u \in V^\natural$ is an element of the appropriate weight. (We will use elements of $V^\natural \otimes V_L$ to denote their equivalence classes in $\mathfrak{m}$.)

We will show that all of the conditions of Theorem 4.1 are satisfied. By considering the weights (with respect to L_0), we see that the abelian subalgebra $\mathfrak{m}_{(0,0)}$ is spanned by elements of the form $1 \otimes \alpha(-1)\iota(1)$ where $\alpha \in \mathscr{L} \otimes_{\mathbb{Z}} \mathbb{R}$. Note that $\mathfrak{m}_{(0,0)}$ is two-dimensional.

By definition $\theta = -1$ on $\mathfrak{m}_{(0,0)}$. Thus, $(\theta(x), \theta(y)) = (x, y)$ for $x, y \in \mathfrak{m}_{(0,0)}$. It also follows from the definition, and symmetry of the form on V_L that

$$\begin{aligned}(\theta(u \otimes e^r), \theta(v \otimes e^{-r})) &= (u \otimes e^{-r}, v \otimes e^r) \\ &= (u \otimes e^r, v \otimes e^{-r})\end{aligned}$$

for $u, v \in V^\natural, r \in \mathscr{L}$.

Consider $-(x, \theta(x))$ for $x \in \mathfrak{m}_r, r \neq 0$. To see that this is strictly positive it is enough to consider elements of the form $x = u \otimes e^r$ where $u \in V^\natural$. Recalling the normalization of the form on V_L,

$$\begin{aligned}(x, \theta(x)) &= (u, u)(-1)^{\langle r, r\rangle/2}(e^r, e^{-r}) \\ &= (-1)^{\langle r, r\rangle}(u, u)(\mathbf{1}, (e^r)_{(-1+\langle r, r\rangle)}e^{-r}) \\ &= (u, u)(\mathbf{1}, \mathbf{1}) < 0.\end{aligned}$$

Therefore, we have the desired properties of the form $(\cdot, \cdot) = (\cdot, \cdot)_{Lie}$, and the contravariant form $(\cdot, \cdot)_0$.

Now if we grade $\mathfrak{m}$, as in [4], by $i = 2m + n \in \mathbb{Z}$ then we see that $\mathfrak{m}$ satisfies the grading condition. Furthermore, condition 3 is satisfied if we take θ to be the involution.

Let $v \in V^\natural$, so that $v \otimes e^r$ is of degree $r = (m, n)$. Then

$$(1 \otimes \alpha(-1)\iota(1))_0 v \otimes e^r = v \otimes \alpha(0)e^r = \langle \alpha, r\rangle v \otimes e^r. \tag{24}$$

Thus, $1 \otimes \alpha(-1)\iota(1)$ acts as the scalar $\langle \alpha, r\rangle$ on $\mathfrak{m}_{(m,n)}$. Thus, all elements of $\mathfrak{m}_{(0,0)}$ act as scalars on the $\mathfrak{m}_{(m,n)}$. As α ranges over $\mathbb{R} \oplus \mathbb{R}$ the action distinguishes between spaces of different degree. This establishes conditions 1, 2 and 3 of Theorem 4.1.

To see that $\mathfrak{m}_{(0,0)} \subset [\mathfrak{m}, \mathfrak{m}]$, let $u, v \in V^\natural_{(2)}$ and $a = e^{(1,1)}$, $b = e^{(1,-1)}$. We will show

$$[u \otimes \iota(a), v \otimes \iota(a^{-1})] \tag{25}$$

and

$$[\iota(b), \iota(b^{-1})] \tag{26}$$

are two linearly independent vectors in $\mathfrak{m}_{(0,0)}$. Since we know that $\mathfrak{m}_{(0,0)}$ is two-dimensional, this will give condition 4 of Theorem 4.1.

By [11, 8.5.44] we have that formula (26) is $\iota(b)_0\iota(b^{-1}) = \bar{b}(-1)\iota(1)$. The formula [11, 8.5.44] also shows $\iota(a)_i\iota(a_{-1}) = 0$ unless $i \leq -3$ and

$$\iota(a)_i\iota(a^{-1}) = \begin{cases} \iota(1) & \text{if } i = -3, \\ \bar{a}(-1)\iota(1) & \text{if } i = -4. \end{cases}$$

Using the Jacobi identity (19) or its component form [11, 8.8.41] and the above, we obtain for formula (25):

$$(u \otimes \iota(a))_0(v \otimes \iota(a^{-1})) = u_3 v \otimes \bar{a}(-1)\iota(1) = c\mathbf{1} \otimes \bar{a}(-1)\iota(1).$$

Since we can pick u and v such that $c \neq 0$ these vectors are linearly independent and we are done.

By definition the radical of the bilinear form on $\mathfrak{m}$ is zero, so by Corollary 4.1, $\mathfrak{m}$ is $\mathfrak{l}/\mathfrak{c}$ for some generalized Kac–Moody algebra $\mathfrak{l}$. In fact $\mathfrak{m} = \mathfrak{g}(M)/\mathfrak{c}$: Because $\mathfrak{m}_{(0,0)}$ is the image of a maximal toral subalgebra, it must also be a maximal toral subalgebra. Define *roots* of $\mathfrak{m}$ as elements $\alpha \in (\mathfrak{m}_{(0,0)})^*$ such that $[h, x] = \alpha(h)x$ for all $x \in \mathfrak{m}$. The grading given by the lattice $\mathscr{L}$ corresponds to the root grading of $\mathfrak{m}$ because we have shown that the elements of $\mathfrak{m}_{(0,0)}$ act as scalars on the $\mathfrak{m}_{(m,n)}$, and that the spaces of different degree are distinguished. It follows from the no-ghost theorem applied to $V^\natural$ that $\mathfrak{m}_{(m,n)} \cong V^\natural_{mn+1}$. We know from [11] that $\dim V^\natural_{mn+1} = c(mn)$. Consider the roots of $\mathfrak{g}(M)$ restricted to $\mathfrak{h}$. Then the dimensions of these restricted root spaces of $\mathfrak{g}(M)$ are given by $c(mn)$. By the specialization (12) of the denominator formula for $\mathfrak{g}(M)$ the generalized Kac–Moody algebra $\mathfrak{g}(M)/\mathfrak{c}$ is isomorphic to $\mathfrak{m}$. □

Since the map given by Corollary 4.1 is an isomorphism on $\mathfrak{n}^\pm$ there are immediate corollaries to Theorem 5.2 and the denominator identity for $\mathfrak{g}(M)$.

Corollary 6.1. *The Monster Lie algebra* $\mathfrak{m}$ *can be written as* $\mathfrak{m} = \mathfrak{u}^+ \oplus \mathfrak{gl}_2 \oplus \mathfrak{u}^-$, *where* $\mathfrak{u}^\pm$ *are free Lie algebras with countably many generators given by Corollary* 5.1.

Corollary 6.2. *The Monster Lie algebra has the denominator formula*

$$u(J(u) - J(v)) = \prod_{\substack{i \in \mathbb{Z}_+ \\ j \in \mathbb{Z}_+ \cup \{-1\}}} (1 - u^i v^j)^{c(ij)}. \tag{27}$$

Appendix. The proof of the no-ghost theorem

In [4] it is shown how to use the no-ghost theorem from string theory to understand some of the structure of the Monster Lie algebra. The proof of that theorem, Theorem 6.2, is reproduced here with the necessary rigor and in a more algebraic context.

The space $V \otimes V_{\mathscr{L}}$ is a vertex algebra with conformal vector. Recall from Section 6.1 that elements of the Virasoro algebra acting on $V \otimes V_{\mathscr{L}}$ satisfy the relations

$$[L_i, L_j] = (i-j)L_{i+j} + \tfrac{26}{12}(i^3 - i)\delta_{i+j,0}.$$

Given a nonzero $r \in \mathscr{L}$, fix nonzero $w \in \mathscr{L}$ such that $\langle w, w\rangle = 0$ and $\langle r, w\rangle \neq 0$. Define operators K_i, $i \in \mathbb{Z}$, on $V \otimes V_{\mathscr{L}}$ by $K_i = (1 \otimes w(-1))_i = 1 \otimes w(i)$.

Let $\mathscr{A}$ be the Lie algebra generated by the operators L_i, K_i with $i \in \mathbb{Z}$. These operators satisfy the relations

$$[L_i, K_j] = -jK_{i+j},$$
$$[K_i, K_j] = 0.$$

The first relation follows from the formula (in $V_{\mathscr{L}}$) $[L_m, w(n)] = -n(w(n+m))$ of [11, 8.7.13] and the fact that $[L_m \otimes 1, 1 \otimes w(n)] = 0$. The second relation holds because $[w(i), w(j)] = \langle w, w\rangle i\delta_{i+j,0}$ [11, 8.6.42] and $\langle w, w\rangle = 0$.

Let $\mathscr{W}$ be the Virasoro subalgebra of $\mathscr{A}$ generated by the $L_i, i \in \mathbb{Z}$, and let $\mathscr{Y}$ be the abelian subalgebra generated by the $K_i, i \in \mathbb{Z}$. Denote by $\mathscr{A}^+$ the subalgebra generated by the L_i, K_i with $i > 0$, let $\mathscr{A}^-$ be the subalgebra generated by the L_i, K_i with $i < 0$, and let $\mathscr{A}^0$ be the subalgebra generated by L_0, K_0. The subalgebras $\mathscr{W}^{\pm}$, $\mathscr{W}^0$ and $\mathscr{Y}^{\pm}$, $\mathscr{Y}^0$ are defined analogously.

The vertex algebra $V \otimes V_{\mathscr{L}}$ is graded by $\mathscr{L}$, because $V_{\mathscr{L}} = S(\hat{\mathfrak{h}}_{\mathbb{Z}}^-) \otimes \mathbb{R}\{\mathscr{L}\}$ has such a grading. The subspace of degree r is $V \otimes S(\hat{\mathfrak{h}}_{\mathbb{Z}}^-) \otimes e^r$. This space will be denoted $\mathscr{H}$. The following subspaces of the $\mathscr{A}$-module $\mathscr{H}$ will be useful:

$$\mathscr{P} = \{v \in \mathscr{H} \mid \mathscr{W}^+ v = 0\},$$
$$\mathscr{T} = \{v \in \mathscr{H} \mid \mathscr{A}^+ v = 0\},$$
$$\mathscr{N} = \text{the radical of the bilinear form } (\cdot,\cdot)_0 \text{on } \mathscr{P},$$
$$\mathscr{K} = U(\mathscr{Y})\mathscr{T}.$$

Denote $V \otimes e^r$ by Ve^r.

Lemma A.1. *With respect to the bilinear form $(\cdot,\cdot)_0$ on $V \otimes V_{\mathscr{L}}$, $L_i^* = L_{-i}$ and $K_i^* = K_{-i}$ for all $i \in \mathbb{Z}$.*

Proof. Let ω be the conformal vector of $V \otimes V_{\mathscr{L}}$, so $L_i = \omega_{i+1}$ and $\theta(\omega) = \omega$. By the definition of the form $(\cdot,\cdot)_0$ and Eq. (21) L_i^* is

$$\mathrm{Res}_{z^{-i-2}} Y(e^{zL_1}(-z^{-2})^{L_0}\omega, z^{-1}).$$

Since $L_1\omega = 0$ and $\mathrm{wt}\,\omega = 2$ we have

$$Y(e^{zL_1}(-z^{-2})^{L_0}\omega, z^{-1}) = \sum_{n\in\mathbb{Z}} \omega_n z^{n-3}.$$

Thus, $L_i^* = \omega_{-i+1} = L_{-i}$.

Now consider $K_i = (1 \otimes w(-1))_i$. By Eq. (21)

$$K_i^* = \mathrm{Res}_{z^{-i-1}} \theta Y(e^{zL_1}(-z^{-2})^{L_0}(1 \otimes w(-1)), z^{-1}).$$

To calculate this, note that $\theta(1 \otimes w(-1)) = -(1 \otimes w(-1))$, that $\mathrm{wt}(1 \otimes w(-1)) = 1$ and that $L_1(1 \otimes w(-1)) = 0$, so

$$-Y(e^{zL_1}(-z^{-2})^{L_0}(1 \otimes w(-1)), z^{-1}) = \sum_{n \in \mathbb{Z}} (1 \otimes w(-1))_n z^{n-1}.$$

We conclude $K_i^* = K_{-i}$. □

Lemma A.2. *The bilinear form $(\cdot,\cdot)_0$ restricted to $\mathscr{H}$ is nondegenerate.*

Proof. The form $(\cdot,\cdot)_0$ on $V \otimes V_{\mathscr{L}}$ is nondegenerate. The form also satisfies $(u,v)_0 = 0$ unless $\deg(u) = \deg(v)$ in $\mathscr{L}$. Thus, the radical of the form $(\cdot,\cdot)_0$ restricted to $\mathscr{H}$ is contained in the radical of the form on $V \otimes V_{\mathscr{L}}$. □

Lemma A.3. $\mathscr{H} = U(\mathscr{A})\mathscr{T}$.

Proof. The bilinear form on $\mathscr{H}$ is nondegenerate (Lemma A.2) and distinct L_0-weight spaces of $\mathscr{H}$ are orthogonal. Thus, the finite-dimensional ith L_0-weight space $\mathscr{H}_i = U(\mathscr{A})\mathscr{T}_i \oplus (U(\mathscr{A})\mathscr{T})_i^{\perp}$. Then there is a decomposition into $\mathscr{A}$-submodules:

$$\mathscr{H} = U(\mathscr{A})\mathscr{T} \oplus (U(\mathscr{A})\mathscr{T})^{\perp}.$$

If the graded submodule $(U(\mathscr{A})\mathscr{T})^{\perp}$ is nonempty then it contains a vector annihilated by $\mathscr{A}^+$ by the following argument: The grading of $\mathscr{H}$ (by weights of L_0) is such that $\mathscr{H} = \coprod_{i \geq (1/2)\langle r,r \rangle} \mathscr{H}_i$. The actions of the generators L_i and K_i, $i > 0$, of $\mathscr{A}^+$ lower the weight of a vector in $\mathscr{H}$. If n is the smallest integer such that $(U(\mathscr{A})\mathscr{T})^{\perp} \cap \mathscr{H}_n$ is nonzero, then this subspace consists of vectors annihilated by $\mathscr{A}^+$. By definition such a vector is in $\mathscr{T}$, hence is in $U(\mathscr{A})\mathscr{T}$, a contradiction. □

Lemma A.4. $\mathscr{K} = \mathscr{T} \oplus \mathrm{rad}(\cdot,\cdot)_0$.

Proof. Note that $\mathscr{Y} = \mathscr{Y}^- \oplus \mathscr{Y}^+ \oplus \mathscr{Y}^0$, so that by the Poincare–Birkhoff–Witt theorem $\mathscr{K} = U(\mathscr{Y}^-)U(\mathscr{Y}^+)U(\mathscr{Y}^0)\mathscr{T}$. By definition of $\mathscr{T}$, $U(\mathscr{Y}^+)\mathscr{T} = \mathscr{T}$ so $\mathscr{K} = U(\mathscr{Y}^-)\mathscr{T}$. Thus,

$$\mathscr{K} = \mathscr{T} \oplus \mathscr{Y}^- U(\mathscr{Y}^-)\mathscr{T}.$$

By Lemma A.1

$$\begin{aligned}(\mathscr{K}, \mathscr{Y}^- U(\mathscr{Y}^-)\mathscr{T})_0 &= (\mathscr{Y}^+\mathscr{K}, U(\mathscr{Y}^-)\mathscr{T})_0 \\ &= 0.\end{aligned}$$

Therefore $\mathscr{Y}^- U(\mathscr{Y}^-)\mathscr{T} \subset \mathrm{rad}(\cdot,\cdot)_0$. Furthermore, $\mathscr{T} \cap \mathrm{rad}(\cdot,\cdot)_0 = 0$ because if $t \in \mathscr{T} \cap \mathrm{rad}(\cdot,\cdot)_0$ then $(t, U(\mathscr{A})\mathscr{T})_0 = (t, U(\mathscr{A}^-)\mathscr{T})_0 = 0$, and by Lemma A.2, $t = 0$. □

Lemma A.5. $\mathscr{K} = Ve^r \oplus rad(\cdot,\cdot)_0$.

Proof. It is immediate from the definition that elements of $\mathscr{K}$ are lowest weight vectors of $\mathscr{Y}$. Furthermore,

$$\begin{aligned}\mathscr{H} &= U(\mathscr{A})\mathscr{T} = [U(\mathscr{W}^-)U(\mathscr{Y}^-)]\mathscr{T} \\ &= \mathscr{W}^- U(\mathscr{W}^-)U(\mathscr{Y}^-)\mathscr{T} \oplus U(\mathscr{Y}^-)\mathscr{T}.\end{aligned}$$

Since no nonzero element of $\mathscr{W}^- U(\mathscr{W}^-)U(\mathscr{Y}^-)\mathscr{T}$ is a lowest weight vector of $\mathscr{Y}$, $\mathscr{K}$ is the subspace of $\mathscr{H}$ of lowest weight vectors of the abelian Lie algebra $\mathscr{Y}$.

In order to describe the lowest weight vectors explicitly consider $S(\hat{\mathfrak{h}}_{\mathbb{Z}}^-)$ as a polynomial algebra on the generators $x_i = w(-i)$, $z_i = r(-i)$, $i > 0$, so that $S(\hat{\mathfrak{h}}_{\mathbb{Z}}^-) = \mathbb{C}[x_i]_{i>0} \otimes \mathbb{C}[z_i]_{i>0}$. The elements $w(i)$, $i \in \mathbb{Z}$ act on $S(\hat{\mathfrak{h}}_{\mathbb{Z}}^-)$ via multiplication, with $w(i) \cdot 1 = 0$ if $i > 0$, $[w(k), w(j)] = 0$ and $[w(k), r(j)] = k\delta_{k+j,0}\langle w, r\rangle$. Thus, the element $w(k)$ acts on $\mathbb{C}[z_i]_{i>0}$ as the differential operator $k\langle r, w\rangle \partial/\partial z_k$ for all $k > 0$.

By definition of the K_i, the lowest weight vectors of the $\mathscr{Y}$-module $\mathscr{H}$ are the lowest weight vectors of the above actions of the $w(i)$, $i \in \mathbb{Z}$, on $\mathbb{C}[x_i]_{i>0} \otimes \mathbb{C}[z_i]_{i>0}$. Since the action of the $w(i)$, $i \in \mathbb{Z}$, commute with the elements $\mathbb{C}[x_i]_{i>0}$, the lowest weight vectors are determined by the elements $q \in \mathbb{C}[z_i]_{i>0}$ satisfying $\partial/\partial z_k q = 0$ for all $k > 0$, so q is a constant. Therefore, the lowest weight vectors of the action of $\mathscr{Y}$ on $S(\hat{\mathfrak{h}}_{\mathbb{Z}}^-)$ correspond to the elements $\mathbb{C}[x_i]_{i>0}$. Thus, $\mathscr{K} = Ve^r \oplus [V \otimes (\mathbb{C}[x_i]_{i>0} \backslash \mathbb{C})e^r]$. Furthermore, $V \otimes (\mathbb{C}[w(-i)]_{i>0} \backslash \mathbb{C})e^r \subset \mathrm{rad}(\cdot,\cdot)_0$, and the form on Ve^r is nondegenerate. □

Let $\mathscr{S} = \mathscr{W}^- U(\mathscr{W}^-)U(\mathscr{Y}^-)\mathscr{T} \subset \mathscr{H}$. This is called the space of "spurious vectors" in [25]. It follows from this definition (and the Poincare–Birkhoff–Witt theorem) that $\mathscr{H} = \mathscr{S} \oplus \mathscr{K}$.

Lemma A.6. *The associative algebra generated by the elements L_i for $i > 0$ is generated by elements mapping $\mathscr{S}_{(1)}$ into $\mathscr{S}$.*

Proof. This is exactly the same argument as in [25]. First we will show that L_1 and $L_2 + \frac{3}{2}L_1^2$ have this property. Any $s \in \mathscr{S}$ can be written

$$s = L_{-1}f_1 + L_{-2}f_2,$$

where $f_1, f_2 \in \mathscr{H}$ since any L_m, $m < 0$ can be written as a polynomial in L_{-1} and L_{-2}. Furthermore, $L_0 s = s$ if and only if

$$L_0L_{-1}f_1 + L_0L_{-2}f_2 = L_{-1}f_1 + L_{-2}f_2,$$

and we may assume $L_0 f_1 = 0, L_0 f_2 = -f_2$.

Thus, $s \in \mathscr{S}_{(1)}$ and $s = L_{-1}f_1 + L_{-2}f_2$ and $L_0 f_2 = -f_2$ and $L_0 f_1 = 0$. Now we compute

$$\begin{aligned}L_1 s &= L_1L_{-1}f_1 + L_1L_{-2}f_2 \\ &= L_{-1}L_1 f_1 + 2L_0 f_1 + 3L_{-1}f_2 + L_{-2}L_1 f_2\end{aligned}$$

and this is in $\mathscr{S}$. Furthermore,

$$\begin{aligned}
&(L_2 + \tfrac{3}{2}L_1L_1)s \\
&\quad = L_2L_{-1}f_1 + \tfrac{3}{2}L_1L_1L_{-1}f_1 + L_2L_{-2}f_2 + \tfrac{3}{2}L_1L_1L_{-2}f_2 \\
&\quad = L_{-1}L_2f_1 + 3L_1f_1 + \tfrac{3}{2}L_1L_{-1}L_1f_1 + L_{-2}L_2f_2 \\
&\qquad + 4L_0f_2 + \tfrac{26}{12}6f_2 + \tfrac{3}{2}L_1L_{-2}L_1f_2 + \tfrac{3^2}{2}L_1L_{-1}f_2 \\
&\quad = L_{-1}\left(L_2 + \tfrac{3}{2}L_1L_1\right)f_1 + L_{-2}\left(L_2 + \tfrac{3}{2}L_1L_1\right)f_2 + 9L_{-1}L_1f_2.
\end{aligned}$$

The above is a spurious vector (it contains L_{-i} with $i > 0$). Note that $D = 26$ is necessary for this computation to work. Since L_1 and $L_2 + \frac{3}{2}L_1^2$ generate the algebra generated by the L_i, where $i > 0$, the lemma is proven. □

Lemma A.7. *$\mathscr{P}_{(1)}$ is the direct sum of $\mathscr{T}_{(1)}$ and $\mathscr{N}_{(1)}$.*

Proof. Let $p \in \mathscr{P}_{(1)}$. Then $p = k + s$ where $k \in \mathscr{K}_{(1)}$ and $s \in \mathscr{S}_{(1)}$; the decomposition is unique. By the preceding lemma a generator u (that is, L_1 or $L_2 + \frac{3}{2}L_1^2$) of $\mathscr{W}^+$ satisfies $0 = up = us + uk \in \mathscr{S} \oplus \mathscr{K}$. Thus, $us = uk = 0$, and we see that s is annihilated by $\mathscr{W}^+$. We conclude that $k \in \mathscr{K} \cap \mathscr{P} = \mathscr{T}$ and $s \in \mathscr{S}_{(1)} \cap \mathscr{P}$. Since $\mathscr{S}$ is orthogonal to $\mathscr{P}$, s must be an element in the radical of the form, $s \in \mathscr{N}_{(1)}$. We conclude that $\mathscr{P}_{(1)} = \mathscr{T}_{(1)} \oplus \mathscr{N}_{(1)}$. □

Theorem 6.2 now follows for $r \neq 0$ because Lemmas A.4 and A.5 imply $Ve^r \approx \mathscr{T}$ so $V_{(1-\langle r,r\rangle/2)e^r} \approx \mathscr{T}_{(1)}$, and Lemma A.7 shows that $\mathscr{T}_{(1)}$ is isomorphic to the quotient $\mathscr{P}_{(1)}/\mathscr{N}_{(1)}$. The isomorphism is naturally a G-isomorphism.

If $r = 0$, first note that

$$(V \otimes V_{\mathscr{L}})_{(1)} = V_{(1)} \oplus (V_{(0)} \otimes (V_{\mathscr{L}})_{(1)}).$$

The subspace $(V_{\mathscr{L}})_{(1)}$ is two-dimensional, spanned by vectors of the form $\alpha(-1)\iota(1)$, $\beta(-1)\iota(1)$ where α, β span $\mathscr{L}$. Furthermore, if $v \in V_{(1)}$ then $L_nv = 0$ for $n > 1$ by the condition that V be $\mathbb{N}$-graded. Since $L_1v \in V_{(0)}$, $L_1v = c1$ for some constant c and since $(L_1v, 1) = (v, L_{-1}1) = 0$ we have $c = 0$. It is easy to show that if $v \in V_{(0)} \otimes (V_{\mathscr{L}})_{(1)}$ then $L_nv = 0$ for $n > 0$. Thus the vectors in $V_{(1)} \oplus (V_{(0)} \otimes (V_{\mathscr{L}})_{(1)})$ are in $\mathscr{P}_{(1)}$, and the radical of the form restricted to this subspace is nondegenerate so Theorem 6.2 follows. □

Acknowledgements

I would like to thank Professors James Lepowsky and Robert L. Wilson for their guidance and many extremely helpful discussions.

References

[1] R. Borcherds, Vertex algebras, Kac–Moody algebras, and the monster, Proc. Natl. Acad. Sci. 83 (1986) 3068–3071.
[2] R. Borcherds, Generalized Kac–Moody Lie algebras, J. Algebra 115 (1988) 501–512.
[3] R. Borcherds, Central extensions of generalized Kac–Moody Lie algebras, J. Algebra 140 (1991) 330–335.
[4] R. Borcherds, Monstrous moonshine and monstrous Lie superalgebras, Invent. Math. 109 (1992) 405–444.
[5] N. Bourbaki, Lie Groups and Lie Algebras, Part 1 (Hermann, Paris, 1975).
[6] J.H. Conway and S.P. Norton, Monstrous moonshine, Bull. London Math. Soc. 11 (1979) 308–339.
[7] C. Dong, Vertex algebras associated with even lattices, J. Algebra 161 (1993) 245–265.
[8] C. Dong and J. Lepowsky, Generalized Vertex Algebras and Relative Vertex Operators, Progress in Math., Vol. 112 (Birkhauser, Boston, 1993).
[9] I. Frenkel, Y.Z. Huang and J. Lepowsky, On axiomatic approaches to vertex operator algebras and modules, preprint, 1989; Mem. Amer. Math. Soc., Vol. 104, No. 494 (American Math. Soc., Providence, RI, 1993).
[10] I. Frenkel, J. Lepowsky and A. Meurman, A natural representation of the Fischer–Griess Monster with the modular function J as character, Proc. Natl. Acad. Sci. USA 81 (1984) 3256–3260.
[11] I. Frenkel, J. Lepowsky and A. Meurman, Vertex Operator Algebras and the Monster (Academic Press, Boston, 1988).
[12] O. Gabber and V. Kac, On defining relations of certain infinite-dimensional Lie algebras, Bull. Amer. Math. Soc. 5 (1981) 185–189.
[13] H. Garland and J. Lepowsky, Lie algebra homology and the MacDonald–Kac formulas, Invent. Math. 34 (1976) 37–76.
[14] R. Gebert and J. Teschner, On the fundamental representation of Borcherds algebras with one imaginary simple root, Lett. Math. Phys., to appear.
[15] P. Goddard and C.B. Thorn, Compatibility of the Dual Pomeron with unitarity and the absence of ghosts in the dual resonance model, Phys. Lett. B 40 (1972) 235–238.
[16] K. Harada, M. Miyamoto and H. Yamada, A generalization of Kac–Moody Lie algebras, to appear.
[17] E. Jurisich, Generalized Kac–Moody algebras and their relation to free Lie algebras, Ph.D. Thesis, Rutgers University, May, 1994.
[18] V. Kac, Infinite-dimensional Lie algebras (Cambridge Univ. Press, Cambridge, 3rd edn., 1990).
[19] S.J. Kang, Generalized Kac–Moody algebras and the modular function j, Math. Ann., to appear.
[20] J. Lepowsky, Lectures on Kac–Moody Lie algebras, Université Paris VI, 1978, unpublished.
[21] H.-S. Li, Equivalence of definitions of vertex algebras, Personal communication, 1991.
[22] H.-S. Li, Symmetric invariant bilinear forms on vertex operator algebras, J. Pure Appl. Algebra 96 (1994) 276–279.
[23] R. Moody, A new class of Lie algebras, J. Algebra 10 (1968) 211–230.
[24] J.P. Serre, A Course in Arithmetic (Springer, New York, 1973).
[25] C. Thorn, A detailed study of the physical state conditions in covariantly quantized string theories, Nucl. Phys. B 286 (1987) 61–77.

Monstrous moonshine

In mathematics, **monstrous moonshine**, or **moonshine theory**, is the unexpected connection between the monster group M and modular functions, in particular, the j function. The term was coined by John Conway and Simon P. Norton in 1979.[1][2][3]

The monstrous moonshine is now known to be underlain by a vertex operator algebra called the moonshine module (or monster vertex algebra) constructed by Igor Frenkel, James Lepowsky, and Arne Meurman in 1988, which has the monster group as its group of symmetries. This vertex operator algebra is commonly interpreted as a structure underlying a two-dimensional conformal field theory, allowing physics to form a bridge between two mathematical areas. The conjectures made by Conway and Norton were proven by Richard Borcherds for the moonshine module in 1992 using the no-ghost theorem from string theory and the theory of vertex operator algebras and generalized Kac–Moody algebras.

Contents

History
The moonshine module
Borcherds' proof
Generalized moonshine
Modular moonshine
Conjectured relationship with quantum gravity
Mathieu moonshine
Origin of the term
Related observations
Notes
Sources
External links

History

In 1978, John McKay found that the first few terms in the Fourier expansion of the normalized J-invariant (sequence A014708 in the OEIS),

$$J(\tau) = \frac{1}{q} + 196884q + 21493760q^2 + 864299970q^3 + 20245856256q^4 + \cdots$$

with $q = e^{2\pi i\tau}$ and τ as the half-period ratio could be expressed in terms of linear combinations of the dimensions of the irreducible representations r_n of the monster group M (sequence A001379 in the OEIS) with small non-negative coefficients. Let r_n = 1, 196883, 21296876, 842609326, 18538750076, 19360062527, 293553734298, ... then,

$$
\begin{aligned}
1 &= r_1 \\
196884 &= r_1 + r_2 \\
21493760 &= r_1 + r_2 + r_3 \\
864299970 &= 2r_1 + 2r_2 + r_3 + r_4 \\
20245856256 &= 3r_1 + 3r_2 + r_3 + 2r_4 + r_5 \\
&= 2r_1 + 3r_2 + 2r_3 + r_4 + r_6 \\
333202640600 &= 5r_1 + 5r_2 + 2r_3 + 3r_4 + 2r_5 + r_7 \\
&= 4r_1 + 5r_2 + 3r_3 + 2r_4 + r_5 + r_6 + r_7
\end{aligned}
$$

(Since there can be several linear relations between the r_n such as $r_1 - r_3 + r_4 + r_5 - r_6 = 0$, the representation may be in more than one way.) McKay viewed this as evidence that there is a naturally occurring infinite-dimensional graded representation of *M*, whose graded dimension is given by the coefficients of *J*, and whose lower-weight pieces decompose into irreducible representations as above. After he informed John G. Thompson of this observation, Thompson suggested that because the graded dimension is just the graded trace of the identity element, the graded traces of nontrivial elements *g* of *M* on such a representation may be interesting as well.

Conway and Norton computed the lower-order terms of such graded traces, now known as McKay–Thompson series T_g, and found that all of them appeared to be the expansions of Hauptmoduln. In other words, if G_g is the subgroup of $\mathrm{SL}_2(\mathbf{R})$ which fixes T_g, then the quotient of the upper half of the complex plane by G_g is a sphere with a finite number of points removed, and furthermore, T_g generates the field of meromorphic functions on this sphere.

Based on their computations, Conway and Norton produced a list of Hauptmoduln, and conjectured the existence of an infinite dimensional graded representation of *M*, whose graded traces T_g are the expansions of precisely the functions on their list.

In 1980, A. Oliver L. Atkin, Paul Fong and Stephen D. Smith produced strong computational evidence that such a graded representation exists, by decomposing a large number of coefficients of *J* into representations of *M*. A graded representation whose graded dimension is *J*, called the moonshine module, was explicitly constructed by Igor Frenkel, James Lepowsky, and Arne Meurman, giving an effective solution to the McKay–Thompson conjecture, and they also determined the graded traces for all elements in the centralizer of an involution of *M*, partially settling the Conway–Norton conjecture. Furthermore, they showed that the vector space they constructed, called the Moonshine Module $V^\natural$, has the additional structure of a vertex operator algebra, whose automorphism group is precisely *M*.

Borcherds proved the Conway–Norton conjecture for the Moonshine Module in 1992. He won the Fields Medal in 1998 in part for his solution of the conjecture.

The moonshine module

The Frenkel–Lepowsky–Meurman construction starts with two main tools:

1. The construction of a lattice vertex operator algebra V_L for an even lattice *L* of rank *n*. In physical terms, this is the chiral algebra for a bosonic string compactified on a torus $\mathbf{R}^n/L$. It can be described roughly as the tensor product of the group ring of *L* with the oscillator representation in *n* dimensions (which is itself isomorphic to a polynomial ring in countably infinitely many generators). For the case in question, one sets *L* to be the Leech lattice, which has rank 24.

2. The orbifold construction. In physical terms, this describes a bosonic string propagating on a quotient orbifold. The construction of Frenkel–Lepowsky–Meurman was the first time orbifolds appeared in conformal field theory. Attached to the –1 involution of the Leech lattice, there is an involution h of V_L, and an irreducible h-twisted V_L-module, which inherits an involution lifting h. To get the Moonshine Module, one takes the fixed point subspace of h in the direct sum of V_L and its twisted module.

Frenkel, Lepowsky, and Meurman then showed that the automorphism group of the moonshine module, as a vertex operator algebra, is M. Furthermore, they determined that the graded traces of elements in the subgroup $2^{1+24}.Co_1$ match the functions predicted by Conway and Norton (Frenkel, Lepowsky & Meurman (1988)).

Borcherds' proof

Richard Borcherds' proof of the conjecture of Conway and Norton can be broken into the following major steps:

1. One begins with a vertex operator algebra V with an invariant bilinear form, an action of M by automorphisms, and with known decomposition of the homogeneous spaces of seven lowest degrees into irreducible M-representations. This was provided by Frenkel–Lepowsky–Meurman's construction and analysis of the Moonshine Module.
2. A Lie algebra $\mathfrak{m}$, called the monster Lie algebra, is constructed from V using a quantization functor. It is a generalized Kac–Moody Lie algebra with a monster action by automorphisms. Using the Goddard–Thorn "no-ghost" theorem from string theory, the root multiplicities are found to be coefficients of J.
3. One uses the Koike–Norton–Zagier infinite product identity to construct a generalized Kac–Moody Lie algebra by generators and relations. The identity is proved using the fact that Hecke operators applied to J yield polynomials in J.
4. By comparing root multiplicities, one finds that the two Lie algebras are isomorphic, and in particular, the Weyl denominator formula for $\mathfrak{m}$ is precisely the Koike–Norton–Zagier identity.
5. Using Lie algebra homology and Adams operations, a twisted denominator identity is given for each element. These identities are related to the McKay–Thompson series T_g in much the same way that the Koike–Norton–Zagier identity is related to J.
6. The twisted denominator identities imply recursion relations on the coefficients of T_g, and unpublished work of Koike showed that Conway and Norton's candidate functions satisfied these recursion relations. These relations are strong enough that one only needs to check that the first seven terms agree with the functions given by Conway and Norton. The lowest terms are given by the decomposition of the seven lowest degree homogeneous spaces given in the first step.

Thus, the proof is completed (Borcherds (1992)). Borcherds was later quoted as saying "I was over the moon when I proved the moonshine conjecture", and "I sometimes wonder if this is the feeling you get when you take certain drugs. I don't actually know, as I have not tested this theory of mine." (Roberts 2009, p. 361)

More recent work has simplified and clarified the last steps of the proof. Jurisich (Jurisich (1998), Jurisich, Lepowsky & Wilson (1995)) found that the homology computation could be substantially shortened by replacing the usual triangular decomposition of the Monster Lie algebra with a decomposition into a sum of gl_2 and two free Lie algebras. Cummins and Gannon showed that the recursion relations automatically imply the McKay Thompson series are either Hauptmoduln or terminate after at most 3 terms, thus eliminating the need for computation at the last step.

Generalized moonshine

Conway and Norton suggested in their 1979 paper that perhaps moonshine is not limited to the monster, but that similar phenomena may be found for other groups.[a] While Conway and Norton's claims were not very specific, computations by Larissa Queen in 1980 strongly suggested that one can construct the expansions of many Hauptmoduln from simple combinations of dimensions of irreducible representations of sporadic groups. In particular, she decomposed the coefficients of McKay-Thompson series into representations of subquotients of the Monster in the following cases:

- T_{2B} and T_{4A} into representations of the Conway group Co_0
- T_{3B} and T_{6B} into representations of the Suzuki group 3.2.*Suz*
- T_{3C} into representations of the Thompson group $Th = F_3$
- T_{5A} into representations of the Harada–Norton group $HN = F_5$
- T_{5B} and T_{10D} into representations of the Hall–Janko group 2.*HJ*
- T_{7A} into representations of the Held group $He = F_7$
- T_{7B} and T_{14C} into representations of $2.A_7$
- T_{11A} into representations of the Mathieu group $2.M_{12}$

Queen found that the traces of non-identity elements also yielded q-expansions of Hauptmoduln, some of which were not McKay–Thompson series from the Monster. In 1987, Norton combined Queen's results with his own computations to formulate the Generalized Moonshine conjecture. This conjecture asserts that there is a rule that assigns to each element g of the monster, a graded vector space $V(g)$, and to each commuting pair of elements (g, h) a holomorphic function $f(g, h, \tau)$ on the upper half-plane, such that:

1. Each $V(g)$ is a graded projective representation of the centralizer of g in M.
2. Each $f(g, h, \tau)$ is either a constant function, or a Hauptmodul.
3. Each $f(g, h, \tau)$ is invariant under simultaneous conjugation of g and h in M, up to a scalar ambiguity.
4. For each (g, h), there is a lift of h to a linear transformation on $V(g)$, such that the expansion of $f(g, h, \tau)$ is given by the graded trace.
5. For any $\begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathbf{SL}_2(\mathbf{Z})$, $f\left(g, h, \frac{a\tau + b}{c\tau + d}\right)$ is proportional to $f\left(g^a h^c, g^b h^d, \tau\right)$.
6. $f(g, h, \tau)$ is proportional to J if and only if $g = h = 1$.

This is a generalization of the Conway–Norton conjecture, because Borcherds's theorem concerns the case where g is set to the identity.

Like the Conway–Norton conjecture, Generalized Moonshine also has an interpretation in physics, proposed by Dixon–Ginsparg–Harvey in 1988 (Dixon, Ginsparg & Harvey (1989)). They interpreted the vector spaces $V(g)$ as twisted sectors of a conformal field theory with monster symmetry, and interpreted the functions $f(g, h, \tau)$ as genus one partition functions, where one forms a torus by gluing along twisted boundary conditions. In mathematical language, the twisted sectors are irreducible twisted modules, and the partition functions are assigned to elliptic curves with principal monster bundles, whose isomorphism type is described by monodromy along a basis of 1-cycles, i.e., a pair of commuting elements.

Modular moonshine

In the early 1990s, the group theorist A. J. E. Ryba discovered remarkable similarities between parts of the character table of the monster, and Brauer characters of certain subgroups. In particular, for an element g of prime order p in the monster, many irreducible characters of an element of order kp whose kth power is g are simple combinations of Brauer characters for an element of order k in the centralizer of g. This was numerical evidence for a phenomenon similar to monstrous moonshine, but for representations in positive characteristic. In particular, Ryba conjectured in 1994 that for each prime factor p in the order of the monster, there exists a graded vertex algebra over the finite field $\mathbf{F}_p$ with an action of the centralizer of an order p element g, such that the graded Brauer character of any p-regular automorphism h is equal to the McKay-Thompson series for gh (Ryba (1996)).

In 1996, Borcherds and Ryba reinterpreted the conjecture as a statement about Tate cohomology of a self-dual integral form of $V^\natural$. This integral form was not known to exist, but they constructed a self-dual form over $\mathbf{Z}[1/2]$, which allowed them to work with odd primes p. The Tate cohomology for an element of prime order naturally has the structure of a super vertex algebra over $\mathbf{F}_p$, and they broke up the problem into an easy step equating graded Brauer super-trace with the McKay-Thompson series, and a hard step showing that Tate cohomology vanishes in odd degree. They proved the vanishing statement for small odd primes, by transferring a vanishing result from the Leech lattice (Borcherds & Ryba (1996)). In 1998, Borcherds showed that vanishing holds for the remaining odd primes, using a combination of Hodge theory and an integral refinement of the no-ghost theorem (Borcherds (1998), Borcherds (1999)).

The case of order 2 requires the existence of a form of $V^\natural$ over a 2-adic ring, i.e., a construction that does not divide by 2, and this was not known to exist at the time. There remain many additional unanswered questions, such as how Ryba's conjecture should generalize to Tate cohomology of composite order elements, and the nature of any connections to generalized moonshine and other moonshine phenomena.

Conjectured relationship with quantum gravity

In 2007, E. Witten suggested that AdS/CFT correspondence yields a duality between pure quantum gravity in (2 + 1)-dimensional anti de Sitter space and extremal holomorphic CFTs. Pure gravity in 2 + 1 dimensions has no local degrees of freedom, but when the cosmological constant is negative, there is nontrivial content in the theory, due to the existence of BTZ black hole solutions. Extremal CFTs, introduced by G. Höhn, are distinguished by a lack of Virasoro primary fields in low energy, and the moonshine module is one example.

Under Witten's proposal (Witten (2007)), gravity in AdS space with maximally negative cosmological constant is AdS/CFT dual to a holomorphic CFT with central charge $c=24$, and the partition function of the CFT is precisely j-744, i.e., the graded character of the moonshine module. By assuming Frenkel-Lepowsky-Meurman's conjecture that moonshine module is the unique holomorphic VOA with central charge 24 and character j-744, Witten concluded that pure gravity with maximally negative cosmological constant is dual to the monster CFT. Part of Witten's proposal is that Virasoro primary fields are dual to black-hole-creating operators, and as a consistency check, he found that in the large-mass limit, the Bekenstein-Hawking semiclassical entropy estimate for a given black hole mass agrees with the logarithm of the corresponding Virasoro primary multiplicity in the moonshine module. In the low-mass regime, there is a small quantum correction to the entropy, e.g., the lowest energy primary fields yield $\ln(196883) \sim 12.19$, while the Bekenstein–Hawking estimate gives $4\pi \sim 12.57$.

Later work has refined Witten's proposal. Witten had speculated that the extremal CFTs with larger cosmological constant may have monster symmetry much like the minimal case, but this was quickly ruled out by independent work of Gaiotto and Höhn. Work by Witten and Maloney (Maloney & Witten (2007)) suggested that pure quantum gravity may not satisfy some consistency checks related to its partition function, unless some subtle properties of complex saddles work out favorably. However, Li–Song–

Strominger (Li, Song & Strominger (2008)) have suggested that a chiral quantum gravity theory proposed by Manschot in 2007 may have better stability properties, while being dual to the chiral part of the monster CFT, i.e., the monster vertex algebra. Duncan–Frenkel (Duncan & Frenkel (2009)) produced additional evidence for this duality by using Rademacher sums to produce the McKay–Thompson series as (2 + 1)-dimensional gravity partition functions by a regularized sum over global torus-isogeny geometries. Furthermore, they conjectured the existence of a family of twisted chiral gravity theories parametrized by elements of the monster, suggesting a connection with generalized moonshine and gravitational instanton sums. At present, all of these ideas are still rather speculative, in part because 3d quantum gravity does not have a rigorous mathematical foundation.

Mathieu moonshine

In 2010, Tohru Eguchi, Hirosi Ooguri, and Yuji Tachikawa observed that the elliptic genus of a K3 surface can be decomposed into characters of the $N = (4,4)$ superconformal algebra, such that the multiplicities of massive states appear to be simple combinations of irreducible representations of the Mathieu group M24.[4] This suggests that there is a sigma-model conformal field theory with K3 target that carries M24 symmetry. However, by the Mukai–Kondo classification, there is no faithful action of this group on any K3 surface by symplectic automorphisms, and by work of Gaberdiel–Hohenegger–Volpato, there is no faithful action on any K3 sigma-model conformal field theory, so the appearance of an action on the underlying Hilbert space is still a mystery.

By analogy with McKay–Thompson series, Cheng suggested that both the multiplicity functions and the graded traces of nontrivial elements of M24 form mock modular forms. In 2012, Gannon proved that all but the first of the multiplicities are non-negative integral combinations of representations of M24, and Gaberdiel–Persson–Ronellenfitsch–Volpato computed all analogues of generalized moonshine functions, strongly suggesting that some analogue of a holomorphic conformal field theory lies behind Mathieu moonshine. Also in 2012, Cheng, Duncan, and Harvey amassed numerical evidence of an umbral moonshine phenomenon where families of mock modular forms appear to be attached to Niemeier lattices. The special case of the A_1^{24} lattice yields Mathieu Moonshine, but in general the phenomenon does not yet have an interpretation in terms of geometry.

Origin of the term

The term "monstrous moonshine" was coined by Conway, who, when told by John McKay in the late 1970s that the coefficient of $\boldsymbol{q}$ (namely 196884) was precisely one more than the degree of the smallest faithful complex representation of the monster group (namely 196883), replied that this was "moonshine" (in the sense of being a crazy or foolish idea).[b] Thus, the term not only refers to the monster group M; it also refers to the perceived craziness of the intricate relationship between M and the theory of modular functions.

Related observations

The monster group was investigated in the 1970s by mathematicians Jean-Pierre Serre, Andrew Ogg and John G. Thompson; they studied the quotient of the hyperbolic plane by subgroups of $SL_2(\mathbf{R})$, particularly, the normalizer $\Gamma_0(p)^+$ of the Hecke congruence subgroup $\Gamma_0(p)$ in $SL(2,\mathbf{R})$. They found that the Riemann surface resulting from taking the quotient of the hyperbolic plane by $\Gamma_0(p)^+$ has genus zero if and only if p is 2, 3, 5, 7, 11, 13, 17, 19, 23, 29, 31, 41, 47, 59 or 71. When Ogg heard about the monster group later on, and noticed that these were precisely the prime factors of the size of M, he published a paper offering a bottle of Jack Daniel's whiskey to anyone who could explain this fact (Ogg (1974)).

Notes

a. Conway, J. and Norton, S. "Monstrous Moonshine", Table 2a, p. 330, http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.103.3704&rep=rep1&type=pdf
b. World Wide Words: Moonshine (http://www.worldwidewords.org/topicalwords/tw-moo1.htm)

Sources

- Borcherds, R. E. (1998), "Modular Moonshine III", *Duke Mathematical Journal*, **93** (1): 129–154, arXiv:math/9801101 (https://arxiv.org/abs/math/9801101), Bibcode:1998math.....1101B (https://ui.adsabs.harvard.edu/abs/1998math.....1101B), doi:10.1215/S0012-7094-98-09305-X (https://doi.org/10.1215%2FS0012-7094-98-09305-X), S2CID 119593942 (https://api.semanticscholar.org/CorpusID:119593942).
- Borcherds, R. E. (1999), "The Fake Monster Formal Group", *Duke Mathematical Journal*, **100** (1): 139–165, arXiv:math/9805123 (https://arxiv.org/abs/math/9805123), doi:10.1215/S0012-7094-99-10005-6 (https://doi.org/10.1215%2FS0012-7094-99-10005-6), S2CID 14404234 (https://api.semanticscholar.org/CorpusID:14404234).
- Borcherds, R. E.; Ryba, A. J. E. (1996), "Modular Moonshine II", *Duke Mathematical Journal*, **83** (2): 435–459, doi:10.1215/S0012-7094-96-08315-5 (https://doi.org/10.1215%2FS0012-7094-96-08315-5), S2CID 119593942 (https://api.semanticscholar.org/CorpusID:119593942).
- Borcherds, Richard (1992), "Monstrous Moonshine and Monstrous Lie Superalgebras" (https://math.berkeley.edu/~reb/papers/monster/monster.pdf) (PDF), *Invent. Math.*, **109**: 405–444, Bibcode:1992InMat.109..405B (https://ui.adsabs.harvard.edu/abs/1992InMat.109..405B), CiteSeerX 10.1.1.165.2714 (https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.165.2714), doi:10.1007/bf01232032 (https://doi.org/10.1007%2Fbf01232032), MR 1172696 (https://www.ams.org/mathscinet-getitem?mr=1172696), S2CID 16145482 (https://api.semanticscholar.org/CorpusID:16145482).
- Conway, John Horton; Norton, Simon P. (1979), "Monstrous Moonshine", *Bull. London Math. Soc.*, **11** (3): 308–339, doi:10.1112/blms/11.3.308 (https://doi.org/10.1112%2Fblms%2F11.3.308), MR 0554399 (https://www.ams.org/mathscinet-getitem?mr=0554399).
- Cummins, C. J.; Gannon, T (1997), "Modular equations and the genus zero property of moonshine functions", *Invent. Math.*, **129** (3): 413–443, Bibcode:1997InMat.129..413C (https://ui.adsabs.harvard.edu/abs/1997InMat.129..413C), doi:10.1007/s002220050167 (https://doi.org/10.1007%2Fs002220050167), S2CID 123073432 (https://api.semanticscholar.org/CorpusID:123073432).
- Dixon, L.; Ginsparg, P.; Harvey, J. (1989), "Beauty and the Beast: superconformal symmetry in a Monster module" (http://projecteuclid.org/euclid.cmp/1104162400), *Comm. Math. Phys.*, **119** (2): 221–241, Bibcode:1988CMaPh.119..221D (https://ui.adsabs.harvard.edu/abs/1988CMaPh.119..221D), doi:10.1007/bf01217740 (https://doi.org/10.1007%2Fbf01217740), S2CID 55102822 (https://api.semanticscholar.org/CorpusID:55102822).
- Du Sautoy, Marcus (2008), *Finding Moonshine, A Mathematician's Journey Through Symmetry*, Fourth Estate, ISBN 978-0-00-721461-7.
- Duncan, John F. R.; Frenkel, Igor B. (2012), *Rademacher sums, moonshine and gravity*, arXiv:0907.4529 (https://arxiv.org/abs/0907.4529), Bibcode:2009arXiv0907.4529D (https://ui.adsabs.harvard.edu/abs/2009arXiv0907.4529D).
- Frenkel, Igor B.; Lepowsky, James; Meurman, Arne (1988), *Vertex Operator Algebras and the Monster*, Pure and Applied Math., vol. 134, Academic Press, ISBN 0-12-267065-5, MR 0996026 (https://www.ams.org/mathscinet-getitem?mr=0996026).

- Gannon, Terry (2000), "Monstrous Moonshine and the Classification of Conformal Field Theories", in Saclioglu, Cihan; Turgut, Teoman; Nutku, Yavuz (eds.), *Conformal Field Theory, New Non-Perturbative Methods in String and Field Theory*, Cambridge Mass: Perseus Publishing, ISBN 0-7382-0204-5 *(Provides introductory reviews to applications in physics)*.
- Gannon, Terry (2006a), "Monstrous Moonshine: The first twenty-five years", *Bull. London Math. Soc.*, **38** (1): 1–33, arXiv:math.QA/0402345 (https://arxiv.org/abs/math.QA/0402345), doi:10.1112/S0024609305018217 (https://doi.org/10.1112%2FS0024609305018217), MR 2201600 (https://www.ams.org/mathscinet-getitem?mr=2201600), S2CID 119626718 (https://api.semanticscholar.org/CorpusID:119626718).
- Gannon, Terry (2006b), *Moonshine beyond the Monster: The Bridge Connecting Algebra, Modular Forms and Physics*, ISBN 978-0-521-83531-2.
- Harada, Koichiro (1999), *Monster*, Iwanami Pub, ISBN 4-00-006055-4 *(The first book about the Monster Group written in Japanese)*.
- Harada, Koichiro (2010), *'Moonshine' of Finite Groups*, European Mathematical Society, ISBN 978-3-03719-090-6, MR 2722318 (https://www.ams.org/mathscinet-getitem?mr=2722318).
- Jurisich, E.; Lepowsky, J.; Wilson, R.L. (1995), "Realizations of the Monster Lie Algebra", *Selecta Math.*, New Series, **1**: 129–161, arXiv:hep-th/9408037 (https://arxiv.org/abs/hep-th/9408037), doi:10.1007/bf01614075 (https://doi.org/10.1007%2Fbf01614075), S2CID 119594012 (https://api.semanticscholar.org/CorpusID:119594012).
- Jurisich, Elizabeth (1998), "Generalized Kac–Moody Lie algebras, free Lie algebras, and the structure of the Monster Lie algebra", *Journal of Pure and Applied Algebra*, **126** (1–3): 233–266, arXiv:1311.3258 (https://arxiv.org/abs/1311.3258), doi:10.1016/s0022-4049(96)00142-9 (https://doi.org/10.1016%2Fs0022-4049%2896%2900142-9), S2CID 119320010 (https://api.semanticscholar.org/CorpusID:119320010).
- Li, Wei; Song, Wei; Strominger, Andrew (21 July 2008), "Chiral gravity in three dimensions", *Journal of High Energy Physics*, **2008** (4): 082, arXiv:0801.4566 (https://arxiv.org/abs/0801.4566), Bibcode:2008JHEP...04..082L (https://ui.adsabs.harvard.edu/abs/2008JHEP...04..082L), doi:10.1088/1126-6708/2008/04/082 (https://doi.org/10.1088%2F1126-6708%2F2008%2F04%2F082), S2CID 2138904 (https://api.semanticscholar.org/CorpusID:2138904).
- Maloney, Alexander; Song, Wei; Strominger, Andrew (2010), "Chiral gravity, log gravity, and extremal CFT", *Phys. Rev. D*, **81** (6): 064007, arXiv:0903.4573 (https://arxiv.org/abs/0903.4573), Bibcode:2010PhRvD..81f4007M (https://ui.adsabs.harvard.edu/abs/2010PhRvD..81f4007M), doi:10.1103/physrevd.81.064007 (https://doi.org/10.1103%2Fphysrevd.81.064007), S2CID 17752143 (https://api.semanticscholar.org/CorpusID:17752143).
- Maloney, Alexander; Witten, Edward (2010), "Quantum Gravity Partition Functions In Three Dimensions", *J. High Energy Phys.*, **2010** (2): 29, arXiv:0712.0155 (https://arxiv.org/abs/0712.0155), Bibcode:2010JHEP...02..029M (https://ui.adsabs.harvard.edu/abs/2010JHEP...02..029M), doi:10.1007/JHEP02(2010)029 (https://doi.org/10.1007%2FJHEP02%282010%2902 9), MR 2672754 (https://www.ams.org/mathscinet-getitem?mr=2672754), S2CID 118363782 (https://api.semanticscholar.org/CorpusID:118363782).
- Ogg, Andrew P. (1974), "Automorphismes de courbes modulaires" (http://archive.numdam.org/ARCHIVE/SDPP/SDPP_1974-1975__16_1/SDPP_1974-1975__16_1_A4_0/SDPP_1974-1975__16_1_A4_0.pdf) (PDF), *Seminaire Delange-Pisot-Poitou. Theorie des nombres, tome 16, no. 1 (1974–1975), exp. no. 7* (in French), vol. 16, pp. 1–8, MR 0417184 (https://www.ams.org/mathscinet-getitem?mr=0417184).
- Roberts, Siobhan (2009), *King of Infinite Space: Donald Coxeter, the Man Who Saved Geometry* (https://books.google.com/books?id=shuJFCWWql4C&pg=PA361), Bloomsbury Publishing USA, p. 361, ISBN 978-080271832-7.

- Ronan, Mark (2006), *Symmetry and the Monster*, Oxford University Press, ISBN 978-0-19-280723-6 *(Concise introduction for the lay reader)*.
- Ryba, A. J. E. (1996), "Modular Moonshine?", in Mason, Geoffrey; Dong, Chongying (eds.), *Moonshine, the Monster, and related topics*, Contemporary Mathematics, vol. 193, Providence, RI: American Mathematical Society, pp. 307–336.
- Witten, Edward (22 June 2007), *Three-Dimensional Gravity Revisited*, arXiv:0706.3359 (https://arxiv.org/abs/0706.3359), Bibcode:2007arXiv0706.3359W (https://ui.adsabs.harvard.edu/abs/2007arXiv0706.3359W).

External links

- Moonshine Bibliography (https://web.archive.org/web/20061205032152/http://cicma.mathstat.concordia.ca/faculty/cummins/moonshine.refs.html) at the Wayback Machine (archived December 5, 2006)

1. A short introduction to Monstrous Moonshine Valdo Tatitscheff January 24, 2019
2. J. Conway and S. Norton. Monstrous Moonshine. Bull. Lond. Math. Soc., 11:308– 339, 1979
3. Mathematicians Chase Moonshine's Shadow Erica Klarreich March 12, 2015 https://www.quantamagazine.org/mathematicians-chase-moonshine-string-theory-connections-20150312/
4. T. Eguchi, H. Ooguri, Y. Tachikawa: Notes on the K3 surface and the Mathieu group M24. Exper. Math. 20 91–96 (2011)

Retrieved from "https://en.wikipedia.org/w/index.php?title=Monstrous_moonshine&oldid=1121164099"

This page was last edited on 10 November 2022, at 21:13 (UTC).

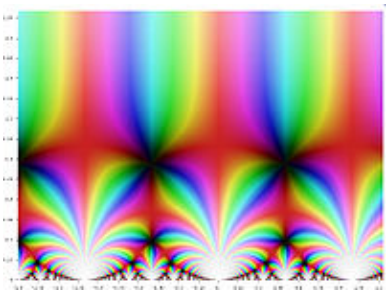
Text is available under the Creative Commons Attribution-ShareAlike License 3.0; additional terms may apply. By using this site, you agree to the Terms of Use and Privacy Policy. Wikipedia® is a registered trademark of the Wikimedia Foundation, Inc., a non-profit organization.

j-invariant

In mathematics, Felix Klein's ***j*-invariant** or ***j* function**, regarded as a function of a complex variable τ, is a modular function of weight zero for $\mathrm{SL}(2, \mathbf{Z})$ defined on the upper half-plane of complex numbers. It is the unique such function which is holomorphic away from a simple pole at the cusp such that

$$j\left(e^{2\pi i/3}\right) = 0, \quad j(i) = 1728 = 12^3.$$

Rational functions of j are modular, and in fact give all modular functions. Classically, the j-invariant was studied as a parameterization of elliptic curves over $\mathbf{C}$, but it also has surprising connections to the symmetries of the Monster group (this connection is referred to as monstrous moonshine).



Klein's j-invariant in the complex plane

Contents

- **Definition**
- **The fundamental region**
- **Class field theory and *j***
- **Transcendence properties**
- **The q-expansion and moonshine**
  - Moonshine
- **Alternate expressions**
- **Expressions in terms of theta functions**
- **Algebraic definition**
- **Inverse function**
- **Pi formulas**
- **Special values**
- **Failure to classify elliptic curves over other fields**
- **References**
  - Notes
  - Other

Definition

The j-invariant can be defined as a function on the upper half-plane $\mathbf{H} = \{\tau \in \mathbf{C},\ \mathrm{Im}(\tau) > 0\}$,

$$j(\tau) = 1728\frac{g_2(\tau)^3}{\Delta(\tau)} = 1728\frac{g_2(\tau)^3}{g_2(\tau)^3 - 27g_3(\tau)^2} = 1728\frac{g_2(\tau)^3}{(2\pi)^{12}\,\eta^{24}(\tau)}$$

with the third definition implying $j(\tau)$ can be expressed as a cube, also since 1728 = 12^3.

The given functions are the modular discriminant $\Delta(\tau) = g_2(\tau)^3 - 27g_3(\tau)^2 = (2\pi)^{12}\,\eta^{24}(\tau)$, Dedekind eta function $\eta(\tau)$, and modular invariants,

$$g_2(\tau) = 60G_4(\tau) = 60 \sum_{(m,n)\neq(0,0)} (m+n\tau)^{-4}$$
$$g_3(\tau) = 140G_6(\tau) = 140 \sum_{(m,n)\neq(0,0)} (m+n\tau)^{-6}$$

where $G_4(\tau)$, $G_6(\tau)$ are Fourier series,

$$G_4(\tau) = \frac{\pi^4}{45}\,E_4(\tau)$$
$$G_6(\tau) = \frac{2\pi^6}{945}\,E_6(\tau)$$

and $E_4(\tau)$, $E_6(\tau)$ are Eisenstein series,

$$E_4(\tau) = 1 + 240\sum_{n=1}^{\infty} \frac{n^3 q^n}{1-q^n}$$
$$E_6(\tau) = 1 - 504\sum_{n=1}^{\infty} \frac{n^5 q^n}{1-q^n}$$

and $q = e^{2\pi i\tau}$ (the square of the nome). The j-invariant can then be directly expressed in terms of the Eisenstein series as,

$$j(\tau) = 1728\frac{E_4(\tau)^3}{E_4(\tau)^3 - E_6(\tau)^2}$$

with no numerical factor other than 1728. This implies a third way to define the modular discriminant,[1]

$$\Delta(\tau) = (2\pi)^{12}\,\frac{E_4(\tau)^3 - E_6(\tau)^2}{1728}$$

For example, using the definitions above and $\tau = 2i$, then the Dedekind eta function $\eta(2i)$ has the exact value,

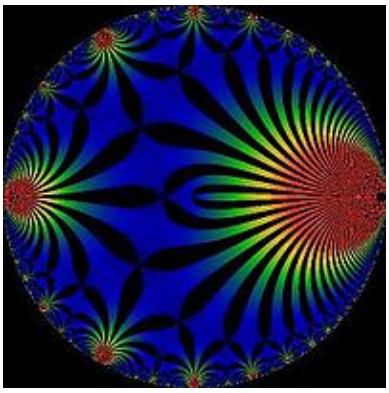
$$\eta(2i) = \frac{\Gamma\left(\frac{1}{4}\right)}{2^{\frac{11}{8}}\pi^{\frac{3}{4}}}$$

implying the transcendental numbers,

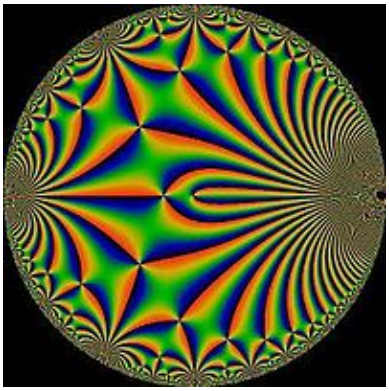
$$g_2(2i) = \frac{11\,\Gamma\left(\frac{1}{4}\right)^8}{2^8\pi^2}, \qquad g_3(2i) = \frac{7\,\Gamma\left(\frac{1}{4}\right)^{12}}{2^{12}\pi^3}$$

but yielding the algebraic number (in fact, an integer),

$$j(2i) = 1728\frac{g_2(2i)^3}{g_2(2i)^3 - 27g_3(2i)^2} = 66^3.$$



Real part of the j-invariant as a function of the square of the nome on the unit disk



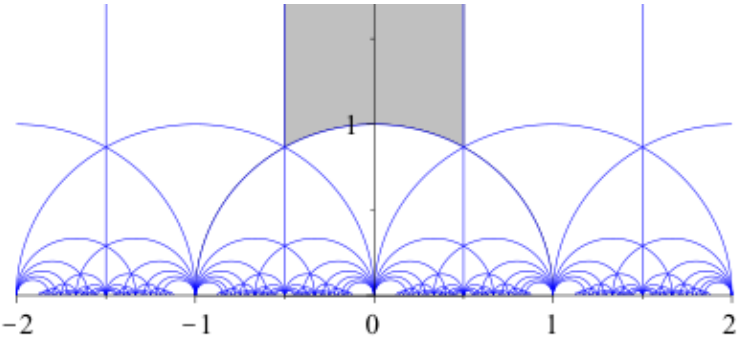
Phase of the j-invariant as a function of the square of the nome on the unit disk

In general, this can be motivated by viewing each τ as representing an isomorphism class of elliptic curves. Every elliptic curve E over $\mathbf{C}$ is a complex torus, and thus can be identified with a rank 2 lattice; that is, a two-dimensional lattice of $\mathbf{C}$. This lattice can be rotated and scaled (operations that preserve the isomorphism class), so that it is generated by 1 and $\tau \in \mathbf{H}$. This lattice corresponds to the elliptic curve $y^2 = 4x^3 - g_2(\tau)x - g_3(\tau)$ (see Weierstrass elliptic functions).

Note that j is defined everywhere in $\mathbf{H}$ as the modular discriminant is non-zero. This is due to the corresponding cubic polynomial having distinct roots.

The fundamental region

It can be shown that Δ is a modular form of weight twelve, and g_2 one of weight four, so that its third power is also of weight twelve. Thus their quotient, and therefore j, is a modular function of weight zero, in particular a holomorphic function $\mathbf{H} \to \mathbf{C}$ invariant under the action of $\mathrm{SL}(2, \mathbf{Z})$. Quotienting out by its centre $\{ \pm \mathrm{I} \}$ yields the modular group, which we may identify with the projective special linear group $\mathrm{PSL}(2, \mathbf{Z})$.



The fundamental domain of the modular group acting on the upper half plane.

By a suitable choice of transformation belonging to this group,

$$\tau \mapsto \frac{a\tau + b}{c\tau + d}, \qquad ad - bc = 1,$$

we may reduce τ to a value giving the same value for j, and lying in the fundamental region for j, which consists of values for τ satisfying the conditions

$$\begin{aligned} &|\tau| \geq 1 \\ &-\tfrac{1}{2} < \Re(\tau) \leq \tfrac{1}{2} \\ &-\tfrac{1}{2} < \Re(\tau) < 0 \Rightarrow |\tau| > 1 \end{aligned}$$

The function $j(\tau)$ when restricted to this region still takes on every value in the complex numbers $\mathbf{C}$ exactly once. In other words, for every c in $\mathbf{C}$, there is a unique τ in the fundamental region such that $c = j(\tau)$. Thus, j has the property of mapping the fundamental region to the entire complex plane.

Additionally two values $\tau,\tau' \in \mathbf{H}$ produce the same elliptic curve iff $\tau = \mathrm{T}(\tau')$ for some $\mathrm{T} \in \mathrm{PSL}(2, \mathbf{Z})$. This means j provides a bijection from the set of elliptic curves over $\mathbf{C}$ to the complex plane.[2]

As a Riemann surface, the fundamental region has genus 0, and every (level one) modular function is a rational function in j; and, conversely, every rational function in j is a modular function. In other words, the field of modular functions is $\mathbf{C}(j)$.

Class field theory and j

The j-invariant has many remarkable properties:

- If τ is any CM point, that is, any element of an imaginary quadratic field with positive imaginary part (so that j is defined), then $j(\tau)$ is an algebraic integer.[3] These special values are called

singular moduli.

- The field extension $\mathbf{Q}[j(\tau),\, \tau]/\mathbf{Q}(\tau)$ is abelian, that is, it has an abelian Galois group.
- Let Λ be the lattice in $\mathbf{C}$ generated by $\{1,\, \tau\}$. It is easy to see that all of the elements of $\mathbf{Q}(\tau)$ which fix Λ under multiplication form a ring with units, called an order. The other lattices with generators $\{1,\, \tau\,'\}$, associated in like manner to the same order define the algebraic conjugates $j(\tau\,')$ of $j(\tau)$ over $\mathbf{Q}(\tau)$. Ordered by inclusion, the unique maximal order in $\mathbf{Q}(\tau)$ is the ring of algebraic integers of $\mathbf{Q}(\tau)$, and values of τ having it as its associated order lead to unramified extensions of $\mathbf{Q}(\tau)$.

These classical results are the starting point for the theory of complex multiplication.

Transcendence properties

In 1937 Theodor Schneider proved the aforementioned result that if τ is a quadratic irrational number in the upper half plane then $j(\tau)$ is an algebraic integer. In addition he proved that if τ is an algebraic number but not imaginary quadratic then $j(\tau)$ is transcendental.

The j function has numerous other transcendental properties. Kurt Mahler conjectured a particular transcendence result that is often referred to as Mahler's conjecture, though it was proved as a corollary of results by Yu. V. Nesterenko and Patrice Phillipon in the 1990s. Mahler's conjecture was that if τ was in the upper half plane then $e^{2\pi i\tau}$ and $j(\tau)$ were never both simultaneously algebraic. Stronger results are now known, for example if $e^{2\pi i\tau}$ is algebraic then the following three numbers are algebraically independent, and thus at least two of them transcendental:

$$j(\tau),\, \frac{j'(\tau)}{\pi},\, \frac{j''(\tau)}{\pi^2}$$

The q-expansion and moonshine

Several remarkable properties of j have to do with its q-expansion (Fourier series expansion), written as a Laurent series in terms of $q = e^{2\pi i\tau}$, which begins:

$$j(\tau) = q^{-1} + 744 + 196884q + 21493760q^2 + 864299970q^3 + 20245856256q^4 + \cdots$$

Note that j has a simple pole at the cusp, so its q-expansion has no terms below q^{-1}.

All the Fourier coefficients are integers, which results in several almost integers, notably Ramanujan's constant:

$$e^{\pi\sqrt{163}} \approx 640320^3 + 744.$$

The asymptotic formula for the coefficient of q^n is given by

$$\frac{e^{4\pi\sqrt{n}}}{\sqrt{2}\, n^{3/4}},$$

as can be proved by the Hardy–Littlewood circle method.[4][5]

Moonshine

More remarkably, the Fourier coefficients for the positive exponents of q are the dimensions of the graded part of an infinite-dimensional graded algebra representation of the monster group called the *moonshine module* – specifically, the coefficient of q^n is the dimension of grade-n part of the moonshine module, the first example being the Griess algebra, which has dimension 196,884, corresponding to the term $196884q$. This startling observation, first made by John McKay, was the starting point for moonshine theory.

The study of the Moonshine conjecture led John Horton Conway and Simon P. Norton to look at the genus-zero modular functions. If they are normalized to have the form

$$q^{-1} + O(q)$$

then John G. Thompson showed that there are only a finite number of such functions (of some finite level), and Chris J. Cummins later showed that there are exactly 6486 of them, 616 of which have integral coefficients.[6]

Alternate expressions

We have

$$j(\tau) = \frac{256(1-x)^3}{x^2}$$

where $x = \lambda(1 - \lambda)$ and λ is the modular lambda function

$$\lambda(\tau) = \frac{\theta_2^4(e^{\pi i \tau})}{\theta_3^4(e^{\pi i \tau})} = k^2(\tau)$$

a ratio of Jacobi theta functions θ_m, and is the square of the elliptic modulus $k(\tau)$.[7] The value of j is unchanged when λ is replaced by any of the six values of the cross-ratio:[8]

$$\left\{\lambda, \frac{1}{1-\lambda}, \frac{\lambda-1}{\lambda}, \frac{1}{\lambda}, \frac{\lambda}{\lambda-1}, 1-\lambda\right\}$$

The branch points of j are at $\{0, 1, \infty\}$, so that j is a Belyi function.[9]

Expressions in terms of theta functions

Define the nome $q = e^{\pi i \tau}$ and the Jacobi theta function,

$$\vartheta(0;\tau) = \vartheta_{00}(0;\tau) = 1 + 2\sum_{n=1}^{\infty}\left(e^{\pi i \tau}\right)^{n^2} = \sum_{n=-\infty}^{\infty} q^{n^2}$$

from which one can derive the auxiliary theta functions. Let,

$$\begin{aligned} a &= \theta_2(q) = \vartheta_{10}(0;\tau) \\ b &= \theta_3(q) = \vartheta_{00}(0;\tau) \\ c &= \theta_4(q) = \vartheta_{01}(0;\tau) \end{aligned}$$

where ϑ_{ij} and θ_n are alternative notations, and $a^4 - b^4 + c^4 = 0$. Then we have the for modular invariants g_2, g_3,

$$g_2(\tau) = \tfrac{2}{3}\pi^4\left(a^8+b^8+c^8\right)$$
$$g_3(\tau) = \tfrac{4}{27}\pi^6\sqrt{\frac{\left(a^8+b^8+c^8\right)^3 - 54(abc)^8}{2}}$$

and modular discriminant,

$$\Delta = g_2^3 - 27g_3^2 = (2\pi)^{12}\left(\tfrac{1}{2}abc\right)^8 = (2\pi)^{12}\eta(\tau)^{24}$$

with Dedekind eta function $\eta(\tau)$. The $j(\tau)$ can then be rapidly computed,

$$j(\tau) = 1728\frac{g_2^3}{g_2^3 - 27g_3^2} = 32\frac{\left(a^8+b^8+c^8\right)^3}{(abc)^8}$$

Algebraic definition

So far we have been considering j as a function of a complex variable. However, as an invariant for isomorphism classes of elliptic curves, it can be defined purely algebraically.[10] Let

$$y^2 + a_1xy + a_3y = x^3 + a_2x^2 + a_4x + a_6$$

be a plane elliptic curve over any field. Then we may perform successive transformations to get the above equation into the standard form $y^2 = 4x^3 - g_2x - g_3$ (note that this transformation can only be made when the characteristic of the field is not equal to 2 or 3). The resulting coefficients are:

$$\begin{aligned} b_2 &= a_1^2 + 4a_2, & b_4 &= a_1a_3 + 2a_4, \\ b_6 &= a_3^2 + 4a_6, & b_8 &= a_1^2a_6 - a_1a_3a_4 + a_2a_3^2 + 4a_2a_6 - a_4^2, \\ c_4 &= b_2^2 - 24b_4, & c_6 &= -b_2^3 + 36b_2b_4 - 216b_6, \end{aligned}$$

where $g_2 = c_4$ and $g_3 = c_6$. We also have the discriminant

$$\Delta = -b_2^2b_8 + 9b_2b_4b_6 - 8b_4^3 - 27b_6^2.$$

The j-invariant for the elliptic curve may now be defined as

$$j = \frac{c_4^3}{\Delta}$$

In the case that the field over which the curve is defined has characteristic different from 2 or 3, this is equal to

$$j = 1728\frac{c_4^3}{c_4^3 - c_6^2}.$$

Inverse function

The inverse function of the j-invariant can be expressed in terms of the hypergeometric function ${}_2F_1$ (see also the article Picard–Fuchs equation). Explicitly, given a number N, to solve the equation $j(\tau) = N$ for τ can be done in at least four ways.

Method 1: Solving the sextic in λ,

$$j(\tau) = \frac{256\big(1-\lambda(1-\lambda)\big)^3}{\big(\lambda(1-\lambda)\big)^2} = \frac{256(1-x)^3}{x^2}$$

where $x = \lambda(1-\lambda)$, and λ is the modular lambda function so the sextic can be solved as a cubic in x. Then,

$$\tau = i\,\frac{{}_2F_1\left(\tfrac{1}{2},\tfrac{1}{2},1;1-\lambda\right)}{{}_2F_1\left(\tfrac{1}{2},\tfrac{1}{2},1;\lambda\right)} = i\frac{\mathrm{M}(1,\sqrt{1-\lambda})}{\mathrm{M}(1,\sqrt{\lambda})}$$

for any of the six values of λ, where M is the arithmetic–geometric mean.[note 1]

Method 2: Solving the quartic in γ,

$$j(\tau) = \frac{27(1+8\gamma)^3}{\gamma(1-\gamma)^3}$$

then for any of the four roots,

$$\tau = \frac{i}{\sqrt{3}}\,\frac{{}_2F_1\left(\tfrac{1}{3},\tfrac{2}{3},1;1-\gamma\right)}{{}_2F_1\left(\tfrac{1}{3},\tfrac{2}{3},1;\gamma\right)}$$

Method 3: Solving the cubic in β,

$$j(\tau) = \frac{64(1+3\beta)^3}{\beta(1-\beta)^2}$$

then for any of the three roots,

$$\tau = \frac{i}{\sqrt{2}}\,\frac{{}_2F_1\left(\tfrac{1}{4},\tfrac{3}{4},1;1-\beta\right)}{{}_2F_1\left(\tfrac{1}{4},\tfrac{3}{4},1;\beta\right)}$$

Method 4: Solving the quadratic in α,

$$j(\tau) = \frac{1728}{4\alpha(1-\alpha)}$$

then,

$$\tau = i\,\frac{{}_2F_1\left(\tfrac{1}{6},\tfrac{5}{6},1;1-\alpha\right)}{{}_2F_1\left(\tfrac{1}{6},\tfrac{5}{6},1;\alpha\right)}$$

One root gives τ, and the other gives $-\frac{1}{\tau}$, but since $j(\tau) = j(-\frac{1}{\tau})$, it makes no difference which α is chosen. The latter three methods can be found in Ramanujan's theory of elliptic functions to alternative bases.

The inversion is applied in high-precision calculations of elliptic function periods even as their ratios become unbounded. A related result is the expressibility via quadratic radicals of the values of j at the points of the imaginary axis whose magnitudes are powers of 2 (thus permitting compass and straightedge constructions). The

latter result is hardly evident since the modular equation for j of order 2 is cubic.[11]

Pi formulas

The Chudnovsky brothers found in 1987,[12]

$$\frac{1}{\pi} = \frac{12}{640320^{3/2}} \sum_{k=0}^{\infty} \frac{(6k)!(163 \cdot 3344418k + 13591409)}{(3k)!(k!)^3(-640320)^{3k}}$$

a proof of which uses the fact that

$$j\left(\frac{1+\sqrt{-163}}{2}\right) = -640320^3.$$

For similar formulas, see the Ramanujan–Sato series.

Special values

The j-invariant vanishes at the "corner" of the fundamental domain:

$$j\left(\tfrac{1+\sqrt{3}i}{2}\right) = 0.$$

Here are a few more special values given in terms of the alternative notation $J(\tau) \equiv \frac{1}{1728}\, j(\tau)$, the first four well known:

$$
\begin{aligned}
J(i) &= J\left(\tfrac{1+i}{2}\right) = 1 \\
J\left(\sqrt{2}i\right) &= \left(\tfrac{5}{3}\right)^3 \\
J(2i) &= \left(\tfrac{11}{2}\right)^3 \\
J\left(2\sqrt{2}i\right) &= \tfrac{125}{216}\left(19+13\sqrt{2}\right)^3 \\
J(4i) &= \tfrac{1}{64}\left(724+513\sqrt{2}\right)^3 \\
J\left(\tfrac{1+2i}{2}\right) &= \tfrac{1}{64}\left(724-513\sqrt{2}\right)^3 \\
J\left(\tfrac{1+2\sqrt{2}i}{3}\right) &= \tfrac{125}{216}\left(19-13\sqrt{2}\right)^3 \\
J(3i) &= \tfrac{1}{27}\left(2+\sqrt{3}\right)^2\left(21+20\sqrt{3}\right)^3 \\
J\left(2\sqrt{3}i\right) &= \tfrac{125}{16}\left(30+17\sqrt{3}\right)^3 \\
J\left(\tfrac{1+7\sqrt{3}i}{2}\right) &= -\tfrac{64000}{7}\left(651+142\sqrt{21}\right)^3 \\
J\left(\tfrac{1+3\sqrt{11}i}{10}\right) &= \tfrac{64}{27}\left(23-4\sqrt{33}\right)^2\left(-77+15\sqrt{33}\right)^3 \\
J\left(\sqrt{21}i\right) &= \tfrac{1}{128}\left(3+\sqrt{7}\right)^5\left(17+7\sqrt{3}+59\sqrt{7}+35\sqrt{21}\right)^3 \\
J\left(\tfrac{\sqrt{30}i}{1}\right) &= \tfrac{1}{4}\left(7+5\sqrt{2}+3\sqrt{5}+2\sqrt{10}\right)^4\left(55+30\sqrt{2}+12\sqrt{5}+10\sqrt{10}\right)^3 \\
J\left(\tfrac{\sqrt{30}i}{2}\right) &= \tfrac{1}{4}\left(7+5\sqrt{2}-3\sqrt{5}-2\sqrt{10}\right)^4\left(55+30\sqrt{2}-12\sqrt{5}-10\sqrt{10}\right)^3 \\
J\left(\tfrac{\sqrt{30}i}{5}\right) &= \tfrac{1}{4}\left(7-5\sqrt{2}+3\sqrt{5}-2\sqrt{10}\right)^4\left(55-30\sqrt{2}+12\sqrt{5}-10\sqrt{10}\right)^3 \\
J\left(\tfrac{\sqrt{30}i}{10}\right) &= \tfrac{1}{4}\left(7-5\sqrt{2}-3\sqrt{5}+2\sqrt{10}\right)^4\left(55-30\sqrt{2}-12\sqrt{5}+10\sqrt{10}\right)^3 \\
J\left(\tfrac{1+\sqrt{31}i}{2}\right) &= \left(1-\frac{1}{64}\left[\left(13+\sqrt{93}\right)\sqrt[3]{4\left(29-3\sqrt{93}\right)}+\left(13-\sqrt{93}\right)\sqrt[3]{4\left(29+3\sqrt{93}\right)}+8\right]^2\right)^3 \\
J\left(\sqrt{70}i\right) &= \left(1+\tfrac{9}{4}\left(303+220\sqrt{2}+139\sqrt{5}+96\sqrt{10}\right)^2\right)^3 \\
J(7i) &= \left(1+\tfrac{9}{32}\sqrt[4]{28}\left(3+\sqrt{7}\right)^3\left(13+3\sqrt{7}+\left(6+\sqrt{7}\right)\sqrt[4]{28}\right)^2\right)^3 \\
J(8i) &= \left(1+\tfrac{9}{4}\sqrt[4]{2}\left(1+\sqrt{2}\right)\left(123+104\sqrt[4]{2}+88\sqrt{2}+73\sqrt[4]{8}\right)^2\right)^3 \\
J(10i) &= \left(1+\tfrac{9}{8}\left(2402+1607\sqrt[4]{5}+1074\sqrt[4]{25}+719\sqrt[4]{125}\right)^2\right)^3 \\
J\left(\tfrac{5i}{2}\right) &= \left(1+\tfrac{9}{8}\left(2402-1607\sqrt[4]{5}+1074\sqrt[4]{25}-719\sqrt[4]{125}\right)^2\right)^3 \\
J\left(2\sqrt{58}i\right) &= \left(1+\tfrac{9}{256}\left(1+\sqrt{2}\right)^5\left(5+\sqrt{29}\right)^5\left(793+907\sqrt{2}+237\sqrt{29}+103\sqrt{58}\right)^2\right)^3 \\
J\left(\tfrac{1+\sqrt{1435}i}{2}\right) &= \left(1-9\left(9892538+4424079\sqrt{5}+1544955\sqrt{41}+690925\sqrt{205}\right)^2\right)^3 \\
J\left(\tfrac{1+\sqrt{1555}i}{2}\right) &= \left(1-9\left(22297077+9971556\sqrt{5}+\left(3571365+1597163\sqrt{5}\right)\sqrt{\tfrac{31+21\sqrt{5}}{2}}\right)^2\right)^3
\end{aligned}
$$

Failure to classify elliptic curves over other fields

The j-invariant is only sensitive to isomorphism classes of elliptic curves over the complex numbers, or more generally, an algebraically closed field. Over other fields there exist examples of elliptic curves whose j-invariant is the same, but are non-isomorphic. For example, let E_1, E_2 be the elliptic curves associated to the polynomials

$$E_1 : y^2 = x^3 - 25x$$
$$E_2 : y^2 = x^3 - 4x,$$

both having j-invariant 1728. Then, the rational points of E_2 can be computed as:

$$E_2(\mathbb{Q}) = \{\infty, (2,0), (-2,0), (0,0)\}$$

since $x^3 - 4x = x(x^2 - 4) = x(x-2)(x+2)$. There are no rational solutions with $y = a \neq 0$. This can be shown using Cardano's formula to show that in that case the solutions to $x^3 - 4x - a^2$ are all irrational. On the other hand, on the set of points

$$\{n(-4,6) : n \in \mathbb{Z}\}$$

the equation for E_2 becomes $36n^2 = -64n^3 + 100n$. Dividing by $4n$ to eliminate the $(0,0)$ solution, the quadratic formula gives the rational solutions:

$$n = \frac{-9 \pm \sqrt{81 - 4 \cdot 16 \cdot (-25)}}{2 \cdot 16} = \frac{-9 \pm 41}{32}.$$

If these curves are considered over $\mathbb{Q}(\sqrt{10})$, there is an isomorphism $E_1(\mathbb{Q}(\sqrt{10})) \cong E_2(\mathbb{Q}(\sqrt{10}))$ sending

$$(x, y) \mapsto (\mu^2 x, \mu^3 y) \text{ where } \mu = \frac{\sqrt{10}}{2}.$$

References

Notes

1. The equality holds if the arithmetic–geometric mean $\mathrm{M}(a, b)$ of complex numbers a, b (such that $a, b \neq 0; a \neq \pm b$) is defined as follows: Let $a_0 = a$, $b_0 = b$, $a_{n+1} = (a_n + b_n)/2$, $b_{n+1} = \pm\sqrt{a_n b_n}$ where the signs are chosen such that $|a_n - b_n| \leq |a_n + b_n|$ for all $n \in \mathbb{N}$. If $|a_n - b_n| = |a_n + b_n|$, the sign is chosen such that $\Im(b_n/a_n) > 0$. Then $\mathrm{M}(a, b) = \lim_{n\to\infty} a_n = \lim_{n\to\infty} b_n$. When a, b are positive real (with $a \neq b$), this definition coincides with the usual definition of the arithmetic–geometric mean for positive real numbers. See The Arithmetic-Geometric Mean of Gauss (https://www.researchgate.net/publication/248675540_The_Arithmetic-Geometric_Mean_of_Gauss) by David A. Cox.

Other

1. Milne, Steven C. (2000). "Hankel Determinants of Eisenstein Series". arXiv:math/0009130v3 (https://arxiv.org/abs/math/0009130v3). The paper uses a non-equivalent definition of Δ, but this has been accounted for in this article.
2. Gareth A. Jones and David Singerman. (1987) Complex functions: an algebraic and geometric viewpoint. Cambridge UP. [1] (https://books.google.com/books?id=jJhWM4vAyVMC)
3. Silverman, Joseph H. (1986). *The Arithmetic of Elliptic Curves*. Graduate Texts in Mathematics. Vol. 106. Springer-Verlag. p. 339. ISBN 978-0-387-96203-0. Zbl 0585.14026 (https://zbmath.org/?format=complete&q=an:0585.14026).
4. Petersson, Hans (1932). "Über die Entwicklungskoeffizienten der automorphen Formen" (https://doi.org/10.1007%2FBF02547776). *Acta Mathematica*. **58** (1): 169–215. doi:10.1007/BF02547776 (https://doi.org/10.1007%2FBF02547776). MR 1555346 (https://mathscinet.ams.org/mathscinet-getitem?mr=1555346).
5. Rademacher, Hans (1938). "The Fourier coefficients of the modular invariant j(τ)". *American Journal of Mathematics*. **60** (2): 501–512. doi:10.2307/2371313 (https://doi.org/10.2307%2F2371313). JSTOR 2371313 (https://www.jstor.org/stable/2371313). MR 1507331 (https://mathscinet.ams.org/mathscinet-getitem?mr=1507331).
6. Cummins, Chris J. (2004). "Congruence subgroups of groups commensurable with *PSL*(2,**Z**)$ of genus 0 and 1" (http://projecteuclid.org/euclid.em/1103749843). *Experimental Mathematics*. **13** (3): 361–382. doi:10.1080/10586458.2004.10504547 (https://doi.org/10.1080%2F10586458.2004.10504547). ISSN 1058-6458 (https://www.worldcat.org/issn/1058-6458). S2CID 10319627 (https://api.semanticscholar.org/CorpusID:10319627). Zbl 1099.11022 (https://zbmath.org/?format=complete&q=an:1099.11022).
7. Chandrasekharan (1985) p.108
8. Chandrasekharan, K. (1985), *Elliptic Functions*, Grundlehren der mathematischen Wissenschaften, vol. 281, Springer-Verlag, p. 110, ISBN 978-3-540-15295-8, Zbl 0575.33001 (https://zbmath.org/?format=complete&q=an:0575.33001)
9. Girondo, Ernesto; González-Diez, Gabino (2012), *Introduction to compact Riemann surfaces and dessins d'enfants*, London Mathematical Society Student Texts, vol. 79, Cambridge: Cambridge University Press, p. 267, ISBN 978-0-521-74022-7, Zbl 1253.30001 (https://zbmath.org/?format=complete&q=an:1253.30001)
10. Lang, Serge (1987). *Elliptic functions*. Graduate Texts in Mathematics. Vol. 112. New-York ect: Springer-Verlag. pp. 299–300. ISBN 978-1-4612-9142-8. Zbl 0615.14018 (https://zbmath.org/?format=complete&q=an:0615.14018).
11. Borwein, Jonathan M.; Borwein, Peter B. (1987). *Pi and the AGM: A Study in Analytic Number Theory and Computational Complexity* (First ed.). Wiley-Interscience. ISBN 0-471-83138-7. Theorem 4.8
12. Chudnovsky, David V.; Chudnovsky, Gregory V. (1989), "The Computation of Classical Constants", *Proceedings of the National Academy of Sciences of the United States of America*, **86** (21): 8178–8182, Bibcode:1989PNAS...86.8178C (https://ui.adsabs.harvard.edu/abs/1989PNAS...86.8178C), doi:10.1073/pnas.86.21.8178 (https://doi.org/10.1073%2Fpnas.86.21.8178), ISSN 0027-8424 (https://www.worldcat.org/issn/0027-8424), JSTOR 34831 (https://www.jstor.org/stable/34831), PMC 298242 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC298242), PMID 16594075 (https://pubmed.ncbi.nlm.nih.gov/16594075).

- Apostol, Tom M. (1976), *Modular functions and Dirichlet Series in Number Theory*, Graduate Texts in Mathematics, vol. 41, New York: Springer-Verlag, MR 0422157 (https://mathscinet.ams.org/mathscinet-getitem?mr=0422157). Provides a very readable introduction and various interesting identities.
  - Apostol, Tom M. (1990), *Modular functions and Dirichlet Series in Number Theory* (https://archive.org/details/modularfunctions0000apos), Graduate Texts in Mathematics, vol. 41 (2nd ed.), doi:10.1007/978-1-4612-0999-7 (https://doi.org/10.1007%2F978-1-4612-0999-7), ISBN 978-0-387-97127-8, MR 1027834 (https://mathscinet.ams.org/mathscinet-getitem?mr=1027834)

- Berndt, Bruce C.; Chan, Heng Huat (1999), "Ramanujan and the modular j-invariant", *Canadian Mathematical Bulletin*, **42** (4): 427–440, doi:10.4153/CMB-1999-050-1 (https://doi.org/10.4153%2FCMB-1999-050-1), MR 1727340 (https://mathscinet.ams.org/mathscinet-getitem?mr=1727340). Provides a variety of interesting algebraic identities, including the inverse as a hypergeometric series.
- Cox, David A. (1989), *Primes of the Form x^2 + ny^2: Fermat, Class Field Theory, and Complex Multiplication*, New York: Wiley-Interscience Publication, John Wiley & Sons Inc., MR 1028322 (https://mathscinet.ams.org/mathscinet-getitem?mr=1028322) Introduces the j-invariant and discusses the related class field theory.
- Conway, John Horton; Norton, Simon (1979), "Monstrous moonshine", *Bulletin of the London Mathematical Society*, **11** (3): 308–339, doi:10.1112/blms/11.3.308 (https://doi.org/10.1112%2Fblms%2F11.3.308), MR 0554399 (https://mathscinet.ams.org/mathscinet-getitem?mr=0554399). Includes a list of the 175 genus-zero modular functions.
- Rankin, Robert A. (1977), *Modular forms and functions*, Cambridge: Cambridge University Press, ISBN 978-0-521-21212-0, MR 0498390 (https://mathscinet.ams.org/mathscinet-getitem?mr=0498390). Provides a short review in the context of modular forms.
- Schneider, Theodor (1937), "Arithmetische Untersuchungen elliptischer Integrale", *Math. Annalen*, **113**: 1–13, doi:10.1007/BF01571618 (https://doi.org/10.1007%2FBF01571618), MR 1513075 (https://mathscinet.ams.org/mathscinet-getitem?mr=1513075), S2CID 121073687 (https://api.semanticscholar.org/CorpusID:121073687).

Retrieved from "https://en.wikipedia.org/w/index.php?title=J-invariant&oldid=1124947639"

This page was last edited on 1 December 2022, at 11:00 (UTC).

Text is available under the Creative Commons Attribution-ShareAlike License 3.0; additional terms may apply. By using this site, you agree to the Terms of Use and Privacy Policy. Wikipedia® is a registered trademark of the Wikimedia Foundation, Inc., a non-profit organization.

arXiv:math/9801101v1 [math.QA] 21 Jan 1998

Modular Moonshine III. *11 October 1994, corrected 18 Nov 1997*

Richard E. Borcherds,*
D.P.M.M.S., 16 Mill Lane, Cambridge, CB2 2SB, England.

Mathematics department, University of California at Berkeley, CA 94720-3840, U. S. A.

reb@dpmms.cam.ac.uk
www.dpmms.cam.ac.uk/~reb

In this paper we complete the proof of Ryba's modular moonshine conjectures [R] that was started in [B-R]. We do this by applying Hodge theory to the cohomology of the monster Lie algebra over the ring of p-adic integers in order to calculate the Tate cohomology groups of elements of the monster acting on the monster vertex algebra.

Contents.

1. Introduction.
Notation.
2. Representations of $\mathbf{Z}_p[G]$.
3. The $\mathbf{Z}[1/2]$-form of the monster Lie algebra.
4. The modular moonshine conjectures for $p \geq 13$.
5. Calculation of some cohomology groups.
6. Elements of type $3B$, $5B$, $7B$, and $13B$.
7. Open problems and conjectures.

1. Introduction.

This paper is a continuation of the earlier paper [B-R] so we will only briefly recall results that are discussed there in more detail. In [B-R] we constructed a modular superalgebra for each element of prime order in the monster, and worked out the structure of this superalgebra for some elements in the monster. In this paper we work out the structure of this superalgebra for the remaining elements of prime order.

Ryba conjectured [R] that for each element of the monster of prime order p of type pA there is a vertex algebra ${}^gV$ defined over the finite field $\mathbf{F}_p$ and acted on by the centralizer $C_M(g)$ of g in the monster group M, with the property that the graded Brauer character $\mathrm{Tr}(h|{}^gV) = \sum_n \mathrm{Tr}(h|{}^gV_n)q^n$ is equal to the Hauptmodul $\mathrm{Tr}(gh|V) = \sum_n \mathrm{Tr}(g|V_n)q^n$ (where V is the graded vertex algebra acted on by the monster constructed by Frenkel, Lepowsky and Meurman [FLM]). In [B-R] the vertex superalgebra ${}^gV$ was defined for any element $g \in M$ of odd prime order to be the sum of the Tate cohomology groups $\hat{H}^0(g, V[1/2]) \oplus \hat{H}^1(g, V[1/2])$ for a suitable $\mathbf{Z}[1/2]$ form $V[1/2]$ of V, and it was shown that $\mathrm{Tr}(h|{}^gV) = \mathrm{Tr}(h|\hat{H}^0(g, V[1/2])) - \mathrm{Tr}(h|\hat{H}^1(g, V[1/2]))$ was equal to the Hauptmodul of $gh \in M$. Hence to prove the modular moonshine conjecture for an element g of type pA it is enough to prove that $\hat{H}^1(g, V[1/2]) = 0$. In [B-R] this was shown by explicit calculation for the elements of type pA for $p \leq 11$, using the fact that these elements commute with an element of type $2B$. For $p \geq 13$ this method does not work as these elements do not commute with an element of type $2B$. The first main theorem of this paper is theorem 4.1 which states that if g is an element of prime order $p \geq 13$ not of type $13B$ then $\hat{H}^1(g, V[1/2]) = 0$ (assuming a certain condition about the monster Lie algebra, whose proof should appear later). Hence Ryba's conjectures are proved for all elements of M of type pA. (Actually this is not quite correct, because the proof for $p = 2$ in [B-R] assumes an unproved technical hypothesis.) The proof we give fails for exactly the cases already proved in [B-R].

We can also ask what happens for the other elements of order less than 13. The cases of elements of types $2B$ or $3C$ are also treated in [B-R]: for type $3C$, $\hat{H}^1(g, V[1/2])$ is again zero, and for type $2B$, $\hat{H}^1(g, V[1/2])$ is zero in even degree and $\hat{H}^0(g, V[1/2])$ is zero in odd degree. This leaves the cases where g is of type $3B$, $5B$, $7B$, or $13B$, when the Tate cohomology groups $\hat{H}^0(g, V[1/2])$ and $\hat{H}^1(g, V[1/2])$ are usually both nonzero in each degree. In these cases the structure of the cohomology groups is determined by our second main result, theorem 6.1, which states that the unique element σ of order 2 in the center of

* Supported by NSF grant DMS-9401186 and by a Royal Society research professorship.

$C_M(g)/O_p(C_M(g))$ acts as 1 on $\hat{H}^0(g, V[1/2])$ and as -1 on $\hat{H}^1(g, V[1/2])$. This determines the modular characters of both cohomology groups because $\mathrm{Tr}(h|\hat{H}^0(g, V[1/2]) + \mathrm{Tr}(h|\hat{H}^1(g, V[1/2]))$ is then given by $\mathrm{Tr}(h\sigma|\hat{H}^0(g, V[1/2])) - \mathrm{Tr}(h\sigma|\hat{H}^1(g, V[1/2]))$ which is the Hauptmodul of the element $gh\sigma \in M$. Hence both $\mathrm{Tr}(h|\hat{H}^0(g, V[1/2]))$ and $\mathrm{Tr}(h|\hat{H}^1(g, V[1/2]))$ can be written explicitly as linear combinations of two Hauptmoduls for any p-regular element $h \in C_M(g)$. We get two new modular vertex superalgebras acted on by double covers of sporadic groups when g has type $3B$ or $5B$: a vertex superalgebra over $\mathbf{F}_3$ acted on by $2.Suz$, and a vertex superalgebra over $\mathbf{F}_5$ acted on by $2.HJ$. For elements of types $7B$ and $13B$ the vertex superalgebras we get are acted on by the double covers $2.A_7$ and $2.A_4$ of alternating groups.

In particular if the integral form V of the monster vertex algebra discussed in [B-R] exists and has the properties conjectured there then all the cohomology groups $\hat{H}^i(g, V)$ are now known for all elements g of prime order in M.

We now discuss the proofs of the two main theorems. We first quickly dispose of the second main theorem: if an element $g \in M$ of odd order commutes with an element of type $2B$ then it is easy to work out its action on the $\mathbf{Z}[1/2]$-form of the monster vertex algebra and hence the Tate cohomology groups $\hat{H}^*(g, V[1/2])$ can be calculated by brute force, which is what we do in sections 5 and 6 for the elements of types $3B$, $5B$, $7B$, and $13B$.

The elements g of prime order $p \geq 13$ do not commute with any elements of type $2B$ (except when g has type $13B$ or $23AB$) so we cannot use the method above. Instead we adapt the proof of the moonshine conjectures in [B]. By using a $\mathbf{Z}_p$-form of the monster Lie algebra rather than a $\mathbf{Q}$-form we can find some complicated relations between the coefficients of the series $\sum_n \mathrm{Tr}(h|\hat{H}^0(g, V[1/2]_n))q^n$ and $\sum_n \mathrm{Tr}(h|\hat{H}^1(g, V[1/2]_n))q^n$. The reason we get more information by using $\mathbf{Z}_p$-forms rather than $\mathbf{Q}_p$-forms is that the group ring $\mathbf{Z}_p[\mathbf{Z}/p\mathbf{Z}]$ has 3 indecomposable modules which are free over $\mathbf{Z}_p$, while the group ring $\mathbf{Q}_p[\mathbf{Z}/p\mathbf{Z}]$ only has 2 indecomposable modules which are free over $\mathbf{Q}_p$, and this extra indecomposable module provides the extra information. The relations between the coefficients we get are similar to the relations defining completely replicable functions, except that some of the relations defining completely replicable functions (about "$(2p-1)/p^2$" of them) are missing. If p is large (≥ 13) we show that these equations have only a finite number of solutions. We then use a computer to find all the solutions, and see that the only solutions imply that $\hat{H}^1(g, V[1/2]) = 0$ for all g of prime order at least 13 other than those of type $13B$. This proof could probably be made to work for some smaller values of p such as 7 and 11, but the difficulty of showing that there are no extra solutions for the equations of the coefficients increases rapidly as p decreases, because for smaller p there are more equations missing.

Y. Martin [Ma] has recently found a conceptual proof that any completely replicable function is a modular function. Cummins and Gannon [CG] have greatly generalized this result using different methods, and showed that the functions are Hauptmoduls. It seems possible that their methods could be extended to replace the computer calculations in section 4, although Cummins has told me that this would probably require adding in some "missing" relations corresponding to the case $(p, mn) \neq 0$ in proposition 3.4.

We summarize the results of this paper and of [B-R] about the vertex superalgebras ${}^gV$ for all elements g of prime order p in the monster. If g is of type pA, or pB for $p > 13$, or $3C$, then ${}^gV$ is just a vertex algebra and not a superalgebra. If $p \geq 13$ this is proved in this paper by an argument which works for any self dual form of the monster vertex algebra but which relies on computer calculations and on a so far unpublished argument about integral forms of the monster Lie algebra. For $p < 13$ (or $p = 23$) this is proved in [B-R] by calculating explicitly for a certain $\mathbf{Z}[1/2]$ form of the monster vertex algebra; this avoids computer calculations but only works for one particular form of the monster vertex algebra. Also, if $p = 2$ the calculation depends on an as yet unproved technical assumption about the Dong-Mason-Montague construction of the monster vertex algebra from an element of type $3B$ ([DM] or [M]). If g is of type pB for $2 \leq p \leq 13$ then the super part of ${}^gV$ does not vanish. If $p = 2$ then ${}^gV$ is calculated explicitly in [B-R] (again using the unproved technical assumption) and it turns out that its ordinary part vanishes in odd degrees and its super part vanishes in even degrees. If $3 \leq p \leq 13$ then ${}^gV$ is calculated in this paper, and we find that there is an element in $C_M(g)$ acting as -1 on the super part and as 1 on the ordinary part of ${}^gV$. In all cases we have explicitly described the modular characters of $C_M(g)$ acting on the ordinary or super part of any degree piece of ${}^gV$ in terms of the coefficients of certain Hauptmoduls.

Notation.

$[A]$ is the element of K represented by the module A.

A, B, C *G-modules.*

A_n *The alternating group on n symbols.*

Aut *The automorphism group of something.*

$c_g^+(n)$ *The n'th coefficient of the Hauptmodul of* $g \in M$*, equal to* $\mathrm{Tr}(1|^g V_n) = \dim(\hat{H}^0(g, V[1/2]_n)) - \dim(\hat{H}^1(g, V[1/2]_n))$.

$c_g^-(n)$ $\dim(\hat{H}^0(g, V[1/2]_n)) + \dim(\hat{H}^1(g, V[1/2]_n))$.

$c_{m,n}$ *Defined in proposition 4.3.*

$\mathbf{C}$ *The complex numbers.*

$C_M(g)$ *The centralizer of g in the group M.*

Co_1 *Conway's largest sporadic simple group.*

E *The positive subalgebra of the monster Lie algebra.*

F *The negative subalgebra of the monster Lie algebra.*

$\mathbf{F}_p$ *The finite field with p elements.*

f *A homomorphism from K to* $\mathbf{Q}$ *defined in lemma 2.3.*

g *An element of G, usually of order p.*

g_i *The element* g^i *of G, used when it is necessary to distinguish the multiplication in the group ring from some other multiplication.*

$\langle g \rangle$ *The group generated by g.*

G *A group, often cyclic of prime order p and generated by g.*

$h(A), h_n$ *See section 5.*

H *The Cartan subalgebra of the monster Lie algebra.*

$\hat{H}^i(G, A)$ *A Tate cohomology group of the finite group G with coefficients in the G-module A.*

$\hat{H}^i(g, A)$ *means* $\hat{H}^i(\langle g \rangle, A)$*, where* $\langle g \rangle$ *is the cyclic group generated by g.*

$\hat{H}^*(g, A)$ *The sum of the Tate cohomology groups* $\hat{H}^0(g, A)$ *and* $\hat{H}^1(g, A)$*, considered as a super module.*

HJ *The Hall-Janko sporadic simple group (sometimes denoted* J_2*).*

I *The indecomposable module of dimension* $p-1$ *over* $\mathbf{Z}_p[G]$*, isomorphic to the kernel of the natural map from* $\mathbf{Z}_p[G]$ *to* $\mathbf{Z}_p$ *and to the quotient* $\mathbf{Z}_p[G]/N_G\mathbf{Z}_p$.

Im *The image of a map.*

K *A ring which is a free* $\mathbf{Q}$*-module with a basis of 3 elements* $[\mathbf{Z}_p] = 1$*,* $[\mathbf{Z}_p[G]]$*, and* $[I]$.

Ker *The kernel of a map.*

$\Lambda, \hat{\Lambda}$ *The Leech lattice and a double cover of the Leech lattice.*

$\Lambda^n(A)$ *The n'th exterior power of A.*

$\Lambda^*(A)$ *The exterior algebra* $\oplus_n \Lambda^n(A)$ *of A.*

L *An even lattice.*

$\mathbf{m}$ *The monster Lie algebra.*

M *The monster simple group.*

N_G *The element* $\sum_{g \in G} g$ *of* $\mathbf{Z}_p[G]$.

$O_p(G)$ *The largest normal p-subgroup of the finite group G.*

p *A prime, usually the order of g.*

$\mathbf{Q}, \mathbf{Q}_p$ *The rational numbers and the field of p-adic numbers.*

ρ *The Weyl vector of the monster Lie algebra.*

$\mathbf{R}$ *The real numbers.*

R_p *A finite extension of the p-adic integers.*

$R(i)$ *Defined just before lemma 3.2.*

σ *An automorphism of type* $2B$ *in the monster or the automorphism* -1 *of the Leech lattice.*

S_n *A symmetric group.*

$S^n(A)$ *The n'th symmetric power of A.*

$S^*(A)$ *The symmetric algebra* $\oplus_n S^n(A)$ *of A.*

Suz *Suzuki's sporadic simple group.*

Tr $\mathrm{Tr}(g|A)$ *is the usual trace of g on a module A if A is a module over a ring of characteristic 0, and the Brauer trace if A is a module over a field of finite characteristic.*

$V[1/n]$ *A* $\mathbf{Z}[1/n]$*-form of the monster vertex algebra.*

V_Λ *The integral form of the vertex algebra of* $\hat{\Lambda}$.

V_n *The degree n piece of V.*
V^n *An eigenspace of some group acting on V.*
${}^gV$ *A modular vertex algebra or superalgebra given by $\hat{H}^*(g, V[1/n])$ for some n coprime to $|g|$.*
$\mathbf{Z}$ *The integers.*
$\mathbf{Z}_p$ *The ring of p-adic integers.*
ω *A cube root of 1 or a conformal vector.*
Ω *The Laplace operator on $\Lambda^*(E)$.*

2. Representations of $\mathbf{Z}_p[G]$.

We give some auxiliary results about modules over $\mathbf{Z}_p[G]$ where G is a cyclic group generated by g of prime order p, and in particular calculate the exterior and symmetric algebras of all indecomposable modules. All modules will be free over $\mathbf{Z}_p$ and will be either finitely generated or $\mathbf{Z}$ graded with finitely generated pieces of each degree. All tensor products will be taken over $\mathbf{Z}_p$.

Recall from [B-R section 2] that there are 3 indecomposable modules over $\mathbf{Z}_p[G]$, which are $\mathbf{Z}_p$, the group ring $\mathbf{Z}_p[G]$, and the module I of $\mathbf{Z}_p[G]$ that is isomorphic to the kernel of the natural map from $\mathbf{Z}_p[G]$ to $\mathbf{Z}_p$ and to the quotient $\mathbf{Z}_p[G]/N_G\mathbf{Z}_p$ (where $N_G = 1 + g + g^2 + \cdots + g^{p-1}$). Their Tate cohomology groups are given by $\hat{H}^0(g, \mathbf{Z}_p) = \mathbf{Z}/p\mathbf{Z}$, $\hat{H}^1(g, \mathbf{Z}_p) = 0$, $\hat{H}^0(g, \mathbf{Z}_p[G]) = 0$, $\hat{H}^1(g, \mathbf{Z}_p[G]) = 0$, $\hat{H}^0(g, I) = 0$, and $\hat{H}^1(g, I) = \mathbf{Z}/p\mathbf{Z}$.

Lemma 2.1. *The tensor products of these modules are given as follows:*

$$\begin{aligned} \mathbf{Z}_p \otimes X &= X \text{ (for any } X\text{)} \\ \mathbf{Z}_p[G] \otimes X &= \mathbf{Z}_p[G]^{\dim(X)} \text{ (the sum of } \dim(X) \text{ copies of } \mathbf{Z}_p[G]\text{)} \\ I \otimes I &= \mathbf{Z}_p[G]^{p-2} \oplus \mathbf{Z}_p. \end{aligned}$$

Proof. The case $\mathbf{Z}_p \otimes X$ is trivial. If X has a basis $x_1, \ldots, x_n$ then for any fixed i the elements $g^j \otimes g^j(x_i)$ $(0 \le j < p)$ form a basis for a submodule of $\mathbf{Z}_p[G] \otimes X$ isomorphic to $\mathbf{Z}_p[G]$, and $\mathbf{Z}_p[G] \otimes X$ is the direct sum of these submodules for all i, which proves the result about $\mathbf{Z}_p[G] \otimes X$. For the case $I \otimes I$ we look at the cohomology sequence of the short exact sequence

$$0 \to I \otimes I \to I \otimes \mathbf{Z}_p[G] \to I \otimes \mathbf{Z}_p \to 0$$

to see that $\hat{H}^0(I \otimes I) = \mathbf{Z}/p\mathbf{Z}$ and $\hat{H}^1(I \otimes I) = 0$, which implies that $I \otimes I$ must be the sum of $\mathbf{Z}_p$ and some copies of $\mathbf{Z}_p[G]$. The number of copies can be worked out by looking at the dimensions of both sides. This proves lemma 2.1.

Corollary 2.2. *If A and B are $\mathbf{Z}_p[G]$-modules then $\hat{H}^*(g, A \otimes B) = \hat{H}^*(g, A) \otimes \hat{H}^*(g, B)$.*

Proof. Check each of the 9 cases when A and B are indecomposable modules using lemma 2.1. This proves corollary 2.2.

We form the ring K which is a free $\mathbf{Q}$-module with a basis of 3 elements $[\mathbf{Z}_p] = 1$, $[\mathbf{Z}_p[G]]$, and $[I]$ corresponding to the indecomposable modules of $\mathbf{Z}_p[G]$, whose addition and multiplication are those induced by tensor products and direct sums of modules. If A is a $\mathbf{Z}_p[G]$-module then we write $[A]$ for the element of K represented by A. Notice that this is not the same as the (tensor product with $\mathbf{Q}$ of the) Grothendieck ring of finitely generated projective $\mathbf{Z}_p[G]$ modules because we do not assume the relations given by nonsplit short exact sequences. The Grothendieck ring (tensored with $\mathbf{Q}$) is a quotient of our ring K by the ideal generated by $[\mathbf{Z}_p] + [I] - [\mathbf{Z}_p[G]]$.

Lemma 2.3. *There are exactly 3 ring homomorphisms* dim, $\mathrm{Tr}(g|.)$, *and f from K to $\mathbf{Q}$ (which correspond to the 3 elements of the spectrum of K). They are given by*

1. $\dim([\mathbf{Z}_p]) = 1$, $\dim([\mathbf{Z}_p[G]]) = p$, $\dim([I]) = p - 1$.
2. $\mathrm{Tr}(g|[\mathbf{Z}_p]) = 1$, $\mathrm{Tr}(g|[\mathbf{Z}_p[G]]) = 0$, $\mathrm{Tr}(g|[I]) = -1$.
3. $f([\mathbf{Z}_p]) = 1$, $f([\mathbf{Z}_p[G]]) = 0$, $f([I]) = 1$.

Proof. From corollary 2.2 we know that the products in the ring K are given by $[\mathbf{Z}_p[G]][\mathbf{Z}_p[G]] = p[\mathbf{Z}_p[G]]$, $[\mathbf{Z}_p[G]][I] = (p-1)[\mathbf{Z}_p[G]]$, and $[I][I] = (p-2)[\mathbf{Z}_p[G]] + 1$. Hence if $x \in \mathbf{Q}$ and $y \in \mathbf{Q}$ are

the images of $[\mathbf{Z}_p[G]]$ and I under some homomorphism they must satisfy $x^2 = px$, $xy = (p-1)x$, and $y^2 = (p-2)x+1$. The only solutions to these equations are $(x,y) = (p, p-1)$, $(0,1)$, or $(0,-1)$. This proves lemma 2.3.

The first two homomorphisms in lemma 2.3 are the homomorphisms which factor through the Grothendieck ring and will not be very interesting to us. We can use them in place of f in the arguments in the rest of the paper, but this only leads to relations between the coefficients of $\sum_n \dim(V_n)q^n$ and $\sum_n \mathrm{Tr}(g|V_n)q^n$, which are essentially the relations already used in [B] to determine these functions. It is the third homomorphism f which detects the integral structure on $V[1/2]$ which will give us enough extra information to prove theorem 4.1.

In general, taking the trace of any fixed element of a group gives a homomorphism from a representation ring of a group to some field, and quite often (e.g., for the complex representation ring of a group) this gives a 1:1 correspondence between homomorphisms and (certain sorts of) conjugacy classes of the group. Lemma 2.3 gives an example of a homomorphism f not constructed in this way.

A permutation module is a module for G with a basis acted on by G. We recall from [B-R] that a module A is a permutation module if and only if $\hat{H}^1(g, A) = 0$, which in turn is equivalent to saying that A is a sum of copies of $\mathbf{Z}_p$ and $\mathbf{Z}_p[G]$.

Lemma 2.4. *If A is a permutation module over any ring for any finite group g then so are the symmetric powers $S^n(A)$. If all elements of G act on a basis of A as products of cycles of odd length then the exterior powers $\Lambda^n(A)$ are also permutation modules.*

Proof. For symmetric powers this is proved in [B-R]. For the proof for exterior powers, let $a_1, \dots, a_n$ be a basis for A acted on by G. Then G acts on the set of all elements of the form $\pm a_{i_1} \wedge \cdots \wedge a_{i_n}$, which consists of all elements of a basis together with their negatives. If the action of G on the basis of A has the property that all elements of G are products of cycles of odd length, then there is no element of G that maps any element of the form $\pm a_{i_1} \wedge \cdots \wedge a_{i_n}$ to its negative. This means that we can find a G-invariant subset of these elements which forms a basis for $\Lambda^n(A)$. This proves lemma 2.4.

Lemma 2.5. *The exterior powers of indecomposable modules are given by*

$$\begin{aligned}
\Lambda^0(\mathbf{Z}_p) &= \mathbf{Z}_p \\
\Lambda^1(\mathbf{Z}_p) &= \mathbf{Z}_p \\
\Lambda^n(\mathbf{Z}_p) &= 0 \text{ if } n \neq 0, 1 \\
\Lambda^0(\mathbf{Z}_p[G]) &= \mathbf{Z}_p \\
\Lambda^p(\mathbf{Z}_p[G]) &= \mathbf{Z}_p \text{ if } p \text{ is odd, and } I \text{ if } p = 2 \\
\Lambda^n(\mathbf{Z}_p[G]) &= \mathbf{Z}_p[G]^{\binom{p-1}{n}} \text{ if } 1 \leq n \leq p-1 \\
\Lambda^n(\mathbf{Z}_p[G]) &= 0 \text{ if } n > p \\
\Lambda^n(I) &= \mathbf{Z}_p \oplus \mathbf{Z}_p[G]^{(\binom{p-1}{n}-1)/p} \text{ if } n \text{ is even and } 0 \leq n < p \\
\Lambda^n(I) &= I \oplus \mathbf{Z}_p[G]^{(\binom{p-1}{n}+1-p)/p} \text{ if } n \text{ is odd and } 0 \leq n < p \\
\Lambda^n(I) &= 0 \text{ if } n \geq p
\end{aligned}$$

Proof. For the exterior powers of $\mathbf{Z}_p[G]$ with p odd we apply lemma 2.4 to see that all exterior powers are sums of $\mathbf{Z}_p$ and $\mathbf{Z}_p[G]$, and the exterior powers can then be identified by checking the dimension and the trace of g. The cohomology groups of the exterior powers of I can then be worked out by induction using the exact sequence

$$0 \to \Lambda^n(I) \to \Lambda^n(\mathbf{Z}_p[G]) \to \Lambda^{n-1}(I) \to 0.$$

The cohomology groups determine how many copies of $\mathbf{Z}_p$ and I occur in the decomposition into indecomposable modules, so we can work out the exterior powers of I from this. The remaining cases of lemma 2.5 are trivial, so this proves lemma 2.5.

These exterior powers induce operators on the ring K which we also denote by Λ^n. For example, we can use these operators to define Adams operations ψ^n on the ring K by

$$\sum_{n \in \mathbf{Z}} (-1)^n \Lambda^n([A])q^n = \exp(-\sum_{n \in \mathbf{Z}} \psi^n([A])q^n/n).$$

Corollary 2.6. *If p is odd then*

$$\sum_{n\in\mathbf{Z}}(-1)^n f([\Lambda^n(\mathbf{Z}_p)])q^n = 1-q$$
$$\sum_{n\in\mathbf{Z}}(-1)^n f([\Lambda^n(\mathbf{Z}_p[G])])q^n = 1-q^p$$
$$\sum_{n\in\mathbf{Z}}(-1)^n f([\Lambda^n(I)])q^n = (1+q^p)/(1+q)$$

Proof. This follows immediately from lemma 2.5 and lemma 2.3.

Lemma 2.7. *The symmetric powers of indecomposable modules are given by*

$$S^n(\mathbf{Z}_p) = \mathbf{Z}_p \text{ if } n \geq 0$$
$$S^n(\mathbf{Z}_p[G]) = \mathbf{Z}_p[G]^{\binom{p+n-1}{n}/p} \text{ if } (p,n)=1$$
$$S^n(\mathbf{Z}_p[G]) = \mathbf{Z}_p[G]^{(\binom{p+n-1}{n}-1)/p} \oplus \mathbf{Z}_p \text{ if } p|n$$
$$S^n(I) = \mathbf{Z}_p \oplus \mathbf{Z}_p[G]^{(\binom{p+n-2}{n}-1)/p} \text{ if } p|n$$
$$S^n(I) = I \oplus \mathbf{Z}_p[G]^{(\binom{p+n-2}{n}+1-p)/p} \text{ if } n \equiv 1 \bmod p$$
$$S^n(I) = \mathbf{Z}_p[G]^{(\binom{p+n-2}{n})/p} \text{ if } n \not\equiv 0,1 \bmod p$$

Proof. For symmetric powers of $\mathbf{Z}_p$ this is trivial and for symmetric powers of $\mathbf{Z}_p[G]$ it can be proved in the same way as in lemma 2.5. If $p=2$ then the case of I is easy to do as I is just one dimensional. For the case of I when p is odd we let N_G be the element $1+g+\cdots+g^{p-1} \in \mathbf{Z}_p[G]$ and consider the quotient of the symmetric algebra $S^*(\mathbf{Z}_p[G])$ by the ideal (N_G). This is the universal commutative algebra generated by $\mathbf{Z}_p[G]/N_G\mathbf{Z}_p$, so it is $S^*(\mathbf{Z}_p[G]/N_G\mathbf{Z}_p)$ which is isomorphic to $S^*(I)$ because $\mathbf{Z}_p[G]/N_G\mathbf{Z}_p$ is isomorphic to I. Therefore there is an exact sequence

$$0 \to S^{n-1}(\mathbf{Z}_p[G]) \to S^n(\mathbf{Z}_p[G]) \to S^n(I) \to 0.$$

Using the known cohomology groups of the symmetric powers of $\mathbf{Z}_p[G]$ and the exact sequence of cohomology groups of this exact sequence gives the cohomology groups of $S^n(I)$ and hence the number of times that $\mathbf{Z}_p$ and I occur in the decomposition of $S^n(I)$. This proves lemma 2.7.

Corollary 2.8. *The cohomology rings of the symmetric algebras of indecomposable modules are give as follows.*

$\hat{H}^*(S^*(\mathbf{Z}_p))$ is a polynomial algebra with one ordinary generator of degree 1.
$\hat{H}^*(S^*(\mathbf{Z}_p[G]))$ is a polynomial algebra with one ordinary generator of degree p.
$\hat{H}^*(S^*(I))$ is the tensor product of a polynomial algebra with one ordinary generator of degree p and an exterior algebra with one super generator of degree 1 if p is odd. (If p is even then $\hat{H}^*(S^*(I))$ is a polynomial algebra on one super generator of degree 1, but we do not need this case.)

Proof. This follows from the calculation of the symmetric powers in lemma 2.7. The product structure on the symmetric algebras can be worked out using the formula $d(ab) = (da)b \pm a(db)$ relating the cup product with the map d from $\hat{H}^i(C)$ to $\hat{H}^{i+1}(A)$ associated to the exact sequence $0 \to A \to B \to C \to 0$. This proves corollary 2.8.

Lemma 2.9. *Suppose that $A_{-1}=0$, $A_0,\ldots,A_n,A_{n+1}=0$ is a sequence of modules and we have maps d from A_i to A_{i+1} and d^* from A_{i+1} to A_i for all i such that $d^2=0$, $d^{*2}=0$, and $dd^*+d^*d=k$ for some fixed integer k coprime to p. Then*

$$[A_0]-[A_1]+[A_2]-\cdots+(-1)^n[A_n]=0.$$

Proof. The fact that k is a unit in $\mathbf{Z}_p$ implies that A_i is the direct sum $dd^*(A_i) \oplus d^*d(A_i)$ and that d is an isomorphism from $d^*d(A_i)$ to $dd^*(A_{i+1})$. The lemma follows from this because the sequence splits as the sum of these isomorphisms. (Notice that the relation between the $[A_i]$'s does not follow just from the exactness of d; for example, look at the short exact sequence $0 \to I \to \mathbf{Z}_p[G] \to \mathbf{Z}_p \to 0$. It is easy to check that in this case there is a map d^* with $d^{*2}=0$ and $dd^*+d^*d=p$, so the condition that $(k,p)=1$ is also necessary.) This proves lemma 2.9.

Lemma 2.10. *Suppose that A is a finitely generated free module over $\mathbf{Z}$ or $\mathbf{Z}_p$ with a symmetric self dual form (,). Then this induces a symmetric self dual form (defined in the proof) on any exterior power $\Lambda^n(A)$.*

Proof. We define the bilinear form on $\Lambda^n(A)$ by

$$(a_1 \wedge \cdots \wedge a_n, b_1 \wedge \cdots \wedge b_n) = \sum_{\sigma \in S_n} \chi(\sigma)(a_1, b_{\sigma(1)}) \cdots (a_n, b_{\sigma(n)})$$

where $\chi(\sigma)$ is 1 or -1 depending on whether σ is an even or odd permutation. This obviously defines a symmetric bilinear form on $\Lambda^n(A)$ with values in $\mathbf{Z}$ or $\mathbf{Z}_p$, and we have to check it is unimodular. The determinant of $\Lambda^n(A)$ is equal to $a \det(A)^b$ for some constants a and b depending on n, but not on A. We can work out a by letting A be the lattice with a base of orthogonal elements $a_1, a_2, \ldots$ with norms ± 1, when we see that $\Lambda^n(A)$ also has an orthogonal base of elements $a_{i_1} \wedge a_{i_2} \cdots$ with $i_1 < i_2 < \cdots$ with norms ± 1, so both A and $\Lambda^n(A)$ have discriminant 1 and so $a = 1$. Hence if A is self dual then $\Lambda^n(A)$ has discriminant 1 and must also be self dual. This proves lemma 2.10. (It is easy to check that the constant b is equal to $\dim(\Lambda^n(A))/\dim(A)$ but we do not need this.)

The analogue of this lemma for symmetric powers is false. The proof breaks down because the element $a_{i_1} a_{i_2} \cdots$ with $i_1 \le i_2 \le \cdots$ does not have norm ± 1 if some of the i_j's are equal. In section 5 we will construct a graded algebra $h(A)$, such that the homogeneous components of $h(A)$ have self dual symmetric bilinear forms if A does, and $h(A) \otimes \mathbf{Q}_p$ is a sum of tensor products of symmetric powers of A.

The following corollary is purely for entertainment and is not used elsewhere in the paper.

Corollary 2.11. *There is a 28-dimensional self dual positive definite lattice with no roots, acted on by the orthogonal group $O_8^+(\mathbf{F}_2)$.*

Proof. By lemma 2.10 the lattice $\Lambda^2(E_8)$ is self dual, and is obviously acted on by the automorphism group of the E_8 lattice modulo -1, which is just $O_8^+(\mathbf{F}_2)$. It is easy to check that this lattice has no roots either by calculation or by observing that there is no suitable root lattice acted on by $O_8^+(\mathbf{F}_2)$. This proves lemma 2.11.

The lattice $\Lambda^2(E_8)$ has theta function

$$\sum_{\lambda \in \Lambda^2(E_8)} q^{\lambda^2/2} = 1 + 2240q^{3/2} + 98280q^{4/2} + 1790208q^{5/2} + 19138560q^{6/2} + \cdots.$$

The vectors of norm 3 are just the 2240 vectors generating one of the 1-dimensional sublattices of the form $\Lambda^2(L)$ where L is one of the 1120 sublattices of E_8 isomorphic to A_2.

3. The $\mathbf{Z}[1/2]$-form of the monster Lie algebra.

We find some relations between the dimensions of the Tate cohomology groups $\hat{H}^i(g, V[1/2])$ by studying the Lie algebra cohomology of the positive subalgebra of a $\mathbf{Z}_p$-form of the monster Lie algebra. In the next section we will use these relations to compute the cohomology groups in the case when g has order at least 13. We recall that $V[1/2]$ is the self dual $\mathbf{Z}[1/2]$-form of Frenkel, Lepowsky, and Meurman's monster vertex algebra [FLM] described in [B-R] (but the arguments in this section work for any self dual $\mathbf{Z}_p$-form of the monster vertex algebra).

This section should really be about the integral form of the monster Lie algebra, but this has not been constructed yet because of the lack of a construction of a self dual integral form of the monster vertex algebra. However, as we are only looking at the action of elements of odd order p in the monster and tensoring everything with $\mathbf{Z}_p$, a $\mathbf{Z}[1/2]$-form is just as good as an integral form.

We briefly recall some properties of the monster Lie algebra $\mathbf{m}$ from [B]. This is a $II_{1,1}$-graded Lie algebra acted on by the monster, where $II_{1,1}$ is the 2-dimensional even self dual Lorentzian lattice whose element (m, n) has norm $-2mn$. The piece of degree (m, n) of $\mathbf{m}$ is isomorphic to V_{mn} if $(m, n) \ne (0, 0)$ and to $\mathbf{R}^2$ if $(m, n) = (0, 0)$. The Lie algebra $\mathbf{m}$ is the sum of 3 subalgebras E, F, and H, where the Cartan subalgebra H is the piece of degree $(0, 0)$ of $\mathbf{m}$, E is the sum of all pieces of degree (m, n) of $\mathbf{m}$ for $m > 0$, and F is the sum of all pieces of degree (m, n) of $\mathbf{m}$ for $m < 0$.

The $\mathbf{Z}[1/2]$ form on V induces a self dual $\mathbf{Z}[1/2]$-form and hence a $\mathbf{Z}_p$-form on $\mathbf{m}$. From now on we will write $\mathbf{m}[1/2]$ for this $\mathbf{Z}[1/2]$-form. In the preprint version of this paper I implicitly assumed that the

degree (m,n) piece of $\mathbf{m}$ was self dual and isomorphic as a module over the monster to V_{mn}. At the last moment I realized that this is not at all clear. The problem is that although the no-ghost theorem gives an isomorphism of both spaces tensored with $\mathbf{Q}$, and both spaces have $\mathbf{Z}[1/2]$-forms, there is no obvious reason why this isomorphism should map one $\mathbf{Z}[1/2]$ form to the other. (In fact for the fake monster Lie algebra the corresponding isomorphism does not preserve the integral forms.) Fortunately in this paper we only need the following weaker statement:

Assumption. *If $m < p$ then the degree (m,n) piece of $\mathbf{m} \otimes \mathbf{Z}_p$ is self dual under the natural bilinear form and isomorphic to $V_{mn} \otimes \mathbf{Z}_p$ as a $\mathbf{Z}_p$ module acted on by the monster.*

I believe I have a proof of this; if all goes well it will appear in a paper provisionally titled "The fake monster formal group". Until this paper appears all the statements in sections 3 and 4 and the modular moonshine conjectures for elements of type pA ($p \geq 13$) should have this assumption added as a hypothesis.

I do not know if the condition $m < p$ in the assumption above is necessary; if it is not, then this condition (usually added in parentheses) can be missed out of some of the following lemmas. It seems very likely that self duality still holds without it, but I am not so sure about the action of the monster being the same if both m and n are divisible by p.

The monster Lie algebra has a Weyl vector ρ, i.e., a vector in the root lattice $II_{1,1}$ such that $(\rho, \alpha) = -(\alpha,\alpha)/2$ for every simple root α. This follows from [B, theorem 7.2] which states that this simple roots are the vectors $(1,n)$ for $n \geq 0$ or $n = -1$, with multiplicities $c(n)$. Hence the Weyl vector is $\rho = (1,0)$.

We recall some facts about Hodge theory applied to Lie algebra cohomology. The self dual symmetric bilinear form on the monster Lie algebra identifies E with the dual space of F, and using the Cartan involution which maps F to E we see that we have a symmetric bilinear form on the Lie algebra E (which is the restriction of the contravariant form of the monster Lie algebra to E) and in particular we can identify E with its dual. The Lie algebra cohomology is the cohomology of an operator d which maps each space $\Lambda^i(E)$ to $\Lambda^{i+1}(E)$. The adjoint of this operator is denoted by d^*, and we define the Laplace operator Ω by $\Omega = dd^* + d^*d$. As for any generalized Kac-Moody algebra with a Weyl vector, the action of the Laplace operator $\Omega = dd^* + d^*d$ on the degree $\alpha = (m,n) \in II_{1,1}$ piece of $\Lambda^*(E)$ is given by multiplication by $(\alpha, \alpha + 2\rho)/2 = (m-1)n$. The Lie algebra cohomology group $H^i(E)$ of E can be identified with the zero eigenspace of Ω on $\Lambda^i(E)$. The Laplace operator restricted to $\Lambda^1(E) = E$ is half the partial Casimir operator Ω_0 defined in [K, section 2.5]; the extra factor of $1/2$ is necessary to make things work well over rings not containing $1/2$, but this is not important in this paper as we work over the ring of p-adic integers for odd p which always contains $1/2$.

Proposition 3.1. *The operators d and d^* act on the $\mathbf{Z}[1/2]$-form $\Lambda^*(E[1/2]) \subset \Lambda^*(E)$ of the exterior algebra (in all degrees (m,n) with $m < p$).*

Proof. The main point is that all homogeneous pieces of the $\mathbf{Z}[1/2]$-form of the exterior algebra are self dual under the contravariant bilinear form. This follows from lemma 2.10 and the fact that each homogeneous component is a finite sum of tensor products of exterior powers of the homogeneous components of $E[1/2]$. The result now follows from the fact that if d is a homomorphism of finite dimensional free $\mathbf{Z}[1/2]$-modules from A to B and A and B have self dual symmetric inner products then the adjoint d^* from B to A is well defined. This proves proposition 3.1.

We let $R(0)$ be the ring of formal Laurent series $K[[r,q]][1/r, 1/q]$ with coefficients in K, and we let $R(1)$ be the space of power series in $R(0)$ only involving monomials $r^m q^n$ with $p|mn$, and we let $R(2)$ be the subring of $R(0)$ of power series only involving powers of r^p and q^p. So we have the inclusions $R(2) \subset R(1) \subset R(0)$ and $R(1)$ is a module over $R(2)$ (but is not a ring itself). (We can also define $R(0)$, $R(1)$, and $R(2)$ for arbitrary generalized Kac-Moody algebras, in particular for the fake monster Lie algebra: $R(0)$ is a completion of the group ring of the root lattice, $R(1)$ is a subspace corresponding to the elements α of the root lattice with $p|(\alpha,\alpha)/2$, and $R(2)$ is the subring corresponding to p times the root lattice.) If A is a $\mathbf{Z}[1/2]$-module then we write $[A]$ for $[A \otimes_{\mathbf{Z}[1/2]} \mathbf{Z}_p]$.

Lemma 3.2. *If $(p, (m-1)n) = 1$ (and $m < p$) then the piece of degree (m,n) of*

$$[\Lambda^*(E[1/2])] = [\Lambda^0(E[1/2])] - [\Lambda^1(E[1/2])] + [\Lambda^2(E[1/2])] - \cdots$$

vanishes. In other words

$$r^{-1}\Lambda^*(\sum_{m>0,n\in\mathbf{Z}} r^m q^n [V_{mn}])$$

lies in $R(1)$.

Proof. The Laplace operator $\Omega = dd^* + d^*d$ acts as multiplication by $(m-1)n$ on the degree (m,n) piece of $\Lambda^*(E[1/2])$, so the result follows from lemma 2.9 if we let A_i be the degree (m,n) piece of $\Lambda^i(E[1/2])$. This proves lemma 3.2.

Lemma 3.3. *If A is a $\mathbf{Z}_p[G]$-module with $\mathrm{Tr}(g|A) = c^+$, $f([A]) = c^-$, (and $m < p$) then*

$$f(\Lambda^*(r^m q^n A)) = (1 - r^m q^n)^{(c^+ + c^-)/2}(1 + r^m q^n)^{(c^+ - c^-)/2} U$$

where U is a unit of the ring $R(2)$.

Proof. It is sufficient to prove this for an indecomposable module A because both sides are multiplicative. We just check each of the three possibilities for A using lemma 2.6.

If $A = \mathbf{Z}_p$ then $c^+ = c^- = 1$ and $f(\Lambda^*(r^m q^n[\mathbf{Z}_p])) = 1 - r^m q^n$, so lemma 3.3 is true in this case.

If $A = \mathbf{Z}_p[G]$ then $c^+ = c^- = 0$ and $f(\Lambda^*(r^m q^n[\mathbf{Z}_p[G]])) = 1 - r^{mp} q^{np}$ which is a unit in $R(2)$, so lemma 3.3 is true in this case.

If $A = I$ then $c^+ = -1$, $c^- = 1$ and

$$f(\Lambda^*(r^m q^n[I])) = (1 + r^{mp} q^{np})/(1 + r^m q^n) = U(1 + r^m q^n)^{-1},$$

so lemma 3.3 is also true in this case. This proves lemma 3.3.

We define integers $c_g^+(n)$ and $c_g^-(n)$ by

$$c_g^+(n) = \mathrm{Tr}(g|V_n) = \dim(\hat{H}^0(g, V[1/2]_n)) - \dim(\hat{H}^1(g, V[1/2]_n))$$
$$c_g^-(n) = f(V[1/2]_n) = \dim(\hat{H}^0(g, V[1/2]_n)) + \dim(\hat{H}^1(g, V[1/2]_n)).$$

We use the numbers $c_g^+(n)$ and $c_g^-(n)$ rather than the apparently simpler numbers $\dim(\hat{H}^0(g, V[1/2]_n))$ and $\dim(\hat{H}^1(g, V[1/2]_n))$ because they turn out to be coefficients of Hauptmoduls, and because $c_g^+(n)$ is already known.

Proposition 3.4. *If $(p, mn) = 1$ (and $m < p$) then the coefficient of $r^m q^n$ in*

$$r^{-1} \prod_{m>0,n\in\mathbf{Z}} (1 - r^m q^n)^{(c_g^-(mn) + c_g^+(mn))/2}(1 + r^m q^n)^{(c_g^+(mn) - c_g^-(mn))/2}$$

vanishes. In other words this power series lies in $R(1)$.

Proof. We apply the homomorphism f to the expression in lemma 3.2 and use lemma 3.3. We find that the expression in this proposition is equal to an element of $R(1)$ times a unit in $R(2)$, and is therefore still an element of $R(1)$. Therefore its coefficients of $r^m q^n$ vanish unless $p|mn$. This proves proposition 3.4.

Proposition 3.4 can be generalized in the obvious way to any generalized Kac-Moody algebra which has a self dual $\mathbf{Z}_p$-form, a Weyl vector, and an integral root lattice.

We can also ask whether or not the coefficients of $r^m q^n$ in proposition 3.4 vanish when $p|mn$. Some numerical calculations suggest that they usually do not.

4. The modular moonshine conjectures for $p \geq 13$.

In this section we prove the following theorem, which completes the proof of the modular moonshine conjectures of [R section 6] (apart from a small technicality in the case $p = 2$.)

Theorem 4.1. *If $g \in M$ is an element of prime order $p \geq 17$ or an element of type $13A$ then $\hat{H}^1(g, V[1/2]) = 0$.*

This implies that ${}^gV = \hat{H}^0(g, V[1/2]) = \hat{H}^*(g, V[1/2])$ is a vertex algebra whose modular character is given by Hauptmoduls, and whose homogeneous components have the characters of [R definition 2].

The proof of this theorem will occupy the rest of this section. We have to show that the numbers $c_g^-(n)$ of proposition 3.4 are equal to the numbers $c_g^+(n)$, because the difference is twice the dimension of $\hat{H}^1(g, V[1/2]_n)$. We start by summarizing what we know about these numbers.

Lemma 4.2.

1. *The numbers $c_g^-(n)$ and $c_g^+(n)$ satisfy the relations given in proposition 3.4.*
2. *The numbers $c_g^-(n)$ are integers, with $c_g^-(n) \equiv c_g^+(n) \bmod 2$.*
3. *$c_g^-(n) \geq |c_g^+(n)|$ and $(p-2)c_g^+(n) + pc_g^-(n) \leq 2c_1^+(n)$, (where $c_1^+(n)$ is the coefficient of q^n in the elliptic modular function).*
4. *The numbers $c_g^+(n)$ are the coefficients of the Hauptmodul of the element $g \in M$.*

Proof. These properties follow easily from proposition 3.4 together with the fact that $\mathbf{Z}_p \otimes V_n$ has dimension $c_1^+(n)$ and is the sum of $(c_g^-(n) + c_g^+(n))/2$ copies of $\mathbf{Z}_p$, $(c_g^-(n) - c_g^+(n))/2$ copies of I, and some copies of $\mathbf{Z}_p[G]$. This proves lemma 4.2.

We will prove theorem 4.1 by showing that if g satisfies the conditions of theorem 4.1 then the conditions 1 to 4 in lemma 4.2 imply that $c_g^-(n) = c_g^+(n)$. We find a finite set of possible solutions of conditions 1, 2, and 4 of lemma 4.2, and find that for each p only one of these satisfies the condition 3. (There are often several solutions not satisfying condition 3.) The proof is just a long messy calculation and the reader should not waste time looking at it.

We first put the relations into a more convenient form.

Proposition 4.3. *There are integers $c_{m,n}$ defined for $m, n > 0$ (and $m < p$) with the following properties.*

1. $$-\log\left(1 + \sum_{m>0,n>0} c_{m,n} r^m q^n\right) \equiv \sum_{m>0,n>0} \sum_{d|(m,n)} \frac{c_g^{(-)^d}(mn/d^2)}{d} r^m q^n \bmod r^p$$
2. *If $(p, mn) = 1$ then $c_{m+1,n} = c_{m,n+1}$.*

Proof. This follows from proposition 3.4 if we multiply both sides by $(r-q)/rq$ and then take the logarithm of both sides. This proves proposition 4.3.

Lemma 4.4. *If $p > 13$ then the values of $c_g^-(n)$ for $n \leq 21$, $c_g^-(2n)$ for $2n \leq 32$, and $c_g^-(36)$ and $c_g^-(45)$ are given by polynomials in the numbers $c_g^-(i)$ for $i = 1, 2, 4, 5$ with coefficients in $\mathbf{Z}_3$.*

Proof. We can evaluate the elements $c(n), n \leq 21$ and $c(2n), n \leq 16$ using the argument for cases 1 and 2 of lemma 4.7 below. We can evaluate $c(36)$ by looking at the coefficient of r^4q^9 in proposition 4.3 and using the fact that we know $c_{m,n}$ for $m \leq 3$, $n \leq 8$. Similarly we can evaluate $c(45)$ by looking at the coefficient of r^5q^9. Notice that the term $(\sum_{m,n} c_{m,n} r^m q^n)^3/3$ has 3's only in the denominators of coefficients of $r^m q^n$ when $3|m$ and $3|n$, so we do not get problems from this for the coefficients of r^4q^9 and r^5q^9 (but we do get problems from this if we try to work out $c_g^-(27)$ using the same method). This proves lemma 4.4.

If $p = 13$ we run into trouble when we try to determine $c_g^-(26)$; this is why the argument in lemma 4.4 does not work in this case. Also the argument breaks down if we try to work out $c(27)$, because when we look at the coefficient of r^3q^9 we get an extra term $c_g^-(3)/3$ which does not have coefficients in $\mathbf{Z}_3$. This is why we use the coefficients $c_g^-(1)$, $c_g^-(2)$, $c_g^-(4)$, and $c_g^-(5)$ rather than $c_g^-(1)$, $c_g^-(2)$, $c_g^-(3)$, and $c_g^-(5)$.

Lemma 4.4 is the reason why our arguments do not work for $p < 13$ (and do not work so smoothly for $p = 13$). As p gets smaller we have fewer relations to work with and it gets more difficult to determine all coefficients in terms of the $c_g^-(m)$'s for small m. It is probably possible to extend lemma 4.4 to cover some smaller primes with a lot of effort. Fortunately it is not necessary to do the cases $p < 13$ because these cases have already been done by explicit calculations in [B-R].

Lemma 4.5. *If $c_g^-(1)$, $c_g^-(2)$, $c_g^-(4)$, and $c_g^-(5)$ are known mod 3^n for some $n \geq 1$, and $p > 13$, and $(c_g^-(1), 3) = 1$ if $p = 13$, then these values of $c_g^-(m)$ are determined mod 3^{n+1}.*

Proof. By lemma 4.4 we see that we can determine the values mod 3^n of $c_g^-(n)$ for $n \leq 16$, $c_g^-(2n)$ for $2n \leq 32$, and $c_g^-(36)$ and $c_g^-(45)$.

But now if we look at the coefficient of $r^3 q^{3m}$ of 3.4 with $m = 1, 2, 4$, or 5, we see that $c_g^-(9m) + c_g^+(2m)/2 + c_g^-(m)/3 = c_g^-(m)^3/3+$ (some polynomial in known $c_g^-(i)$'s with coefficients that are 3-adic integers). But $c_g^-(m)^3/3 \bmod 3^n$ depends only on $c_g^-(m) \bmod 3^n$ if $n \geq 1$, so we can determine $c_g^-(m)/3 \bmod 3^n$ and hence $c_g^-(m) \bmod 3^{n+1}$. This proves lemma 4.5.

This lemma is the reason that we use 3-adic rather than 2-adic approximation. If we try to prove the lemma above for 2^{n+1} instead of 3^{n+1}, all we find is that we can determine the $c_g^+(m)$'s mod 2^{n+1}, which is useless because we already know these numbers.

Proposition 4.6. *If the numbers $c_g^-(n)$ satisfy the conditions 1, 2, and 4 of lemma 4.2 then the numbers $c_g^-(1)$, $c_g^-(2)$, $c_g^-(4)$, and $c_g^-(5)$ are congruent mod 3^{29} to the coefficients of q, q^2, q^4, and q^5 of one of the following power series. (The second column is a genus zero group whose Hauptmodul appears to be the function with coefficients $c_g^-(n)$. This has not been checked rigorously as it is not necessary for the proof of theorem 4.1.)*

$p = 71$	$\Gamma_0(71)+$	q^{-1}	$+q$	$+q^2$	$+q^3$	$+q^4$	$+2q^5$	$+\cdots$
$p = 59$	$\Gamma_0(59)+$	q^{-1}	$+q$	$+q^2$	$+2q^3$	$+2q^4$	$+3q^5$	$+\cdots$
$p = 47$	$\Gamma_0(47)+$	q^{-1}	$+q$	$+2q^2$	$+3q^3$	$+3q^4$	$+5q^5$	$+\cdots$
$p = 47$	$\Gamma_0(94)+$	q^{-1}	$+q$		$+q^3$	$+q^4$	$+q^5$	$+\cdots$
$p = 41$	$\Gamma_0(41)+$	q^{-1}	$+2q$	$+2q^2$	$+3q^3$	$+4q^4$	$+7q^5$	$+\cdots$
$p = 41$	$\Gamma_0(82\|2)+$	q^{-1}			$+q^3$		$+q^5$	$+\cdots$
$p = 31$	$\Gamma_0(31)+$	q^{-1}	$+3q$	$+3q^2$	$+6q^3$	$+9q^4$	$+13q^5$	$+\cdots$
$p = 31$	$\Gamma_0(62)+$	q^{-1}	$+q$	$+q^2$	$+2q^3$	$+q^4$	$+3q^5$	$+\cdots$
$p = 29$	$\Gamma_0(29)+$	q^{-1}	$+3q$	$+4q^2$	$+7q^3$	$+10q^4$	$+17q^5$	$+\cdots$
$p = 29$	$\Gamma_0(58\|2)+$	q^{-1}	$+q$		$+q^3$		$+q^5$	$+\cdots$
$p = 23$	$\Gamma_0(23)+$	q^{-1}	$+4q$	$+7q^2$	$+13q^3$	$+19q^4$	$+33q^5$	$+\cdots$
$p = 23$	$\Gamma_0(46) + 23$	q^{-1}		$-q^2$	$+q^3$	$-q^4$	$+q^5$	$+\cdots$
$p = 23$	$\Gamma_0(46)+$	q^{-1}	$+2q$	$+q^2$	$+3q^3$	$+3q^4$	$+5q^5$	$+\cdots$
$p = 19$	$\Gamma_0(19)+$	q^{-1}	$+6q$	$+10q^2$	$+21q^3$	$+36q^4$	$+61q^5$	$+\cdots$
$p = 19$	$\Gamma_0(38)+$	q^{-1}	$+2q$	$+2q^2$	$+5q^3$	$+4q^4$	$+9q^5$	$+\cdots$
$p = 19$	$\Gamma_0(38\|2)+$	q^{-1}	$+2q$		$+q^3$		$+3q^5$	$+\cdots$
$p = 17$	$\Gamma_0(17)+$	q^{-1}	$+7q$	$+14q^2$	$+29q^3$	$+50q^4$	$+92q^5$	$+\cdots$
$p = 17$	$\Gamma_0(34)+$	q^{-1}	$+3q$	$+2q^2$	$+5q^3$	$+6q^4$	$+12q^5$	$+\cdots$
$p = 17$	$\Gamma_0(34\|2)+$	q^{-1}	$+q$		$+3q^3$		$+4q^5$	$+\cdots$
$p = 13$	$\Gamma_0(13)+$	q^{-1}	$+12q$	$+28q^2$	$+66q^3$	$+132q^4$	$+258q^5$	$+\cdots$
$p = 13$	$\Gamma_0(26)+$	q^{-1}	$+4q$	$+4q^2$	$+10q^3$	$+12q^4$	$+26q^5$	$+\cdots$
$p = 13$	$\Gamma_0(26\|2)+$	q^{-1}	$+2q$		$+4q^3$		$+6q^5$	$+\cdots$

It is also possible (but unlikely) that there are other solutions for $p = 13$ for which $c_g^-(1)$ does not satisfy the inequalities $0 \leq c_g^-(1) < 3^{10}$.

Proof. For each prime p we use a computer to test all 81 possibilities for $c_g^-(1)$, $c_g^-(2)$, $c_g^-(4)$, and $c_g^-(5)$ mod 3. If $p > 13$ then we can calculate the p-adic expansion of all the coefficients $c_g^-(n)$ recursively using lemma 4.5 above, and we reach a contradiction by looking at coefficients of proposition 4.3 except in the cases above. For $p = 13$ this does not quite work as we have not shown the values mod 3^n determine those mod 3^{n+1}, so we can adopt the crude procedure of just testing all 3^4 possibilities for the $c_g^-(i)$'s mod 3^{n+1} for each solution mod 3^n we have found, and checking to see which of them leads to contradictions. There were some cases for $p = 13$ where this did not lead to a contradiction but did at least lead to the conclusion that $c_g^-(1) < 0$ or $c_g^-(1) \geq 3^{10}$. (If $c_g^-(1)$ is not divisible by 3 then even when $p = 13$ the numbers $c_g^-(i) \bmod 3^n$ determine the numbers $c_g^-(n) \bmod 3^{n+1}$. When $c_g^-(1)$ was divisible by 3, there were several cases where the coefficient $c_g^-(5)$ was not determined mod 3^{n+1} by the identities used by the computer

program. However in all the cases looked at by the computer with $3|c_g^-(1)$, the values of $c_g^-(1)$, $c_g^-(2)$, $c_g^-(4)$, and $3c_g^-(5)$ mod 3^n uniquely determined their values mod 3^{n+1}.) This proves proposition 4.6, at least if one believes the computer calculations. (Anyone who does not like computer calculations in a proof is welcome to redo the calculations by hand.)

Lemma 4.7. *If the coefficients $c_g^-(1)$, $c_g^-(2)$, $c_g^-(4)$, and $c_g^-(5)$ are equal to the coefficients $c_g^+(1)$, $c_g^+(2)$, $c_g^+(4)$, and $c_g^+(5)$ for g an element of prime order $p > 13$ or an element of type $13A$ then all the coefficients are determined by the relations of proposition 3.4.*

Proof. This proof is just a long case by case check using induction. We will repeatedly use the fact that the coefficients of $r^m q^n$ of both sides of proposition 4.3 are equal.

Case 1. $c_g^-(2n)$ for $2n \not\equiv 0 \bmod p$. By looking at the coefficient of r^2q^n and using the fact that $c_{2,n} = c_{1,n+1}$ (as $(p,n) = 1$) we see that $c_g^-(2n)$ can be written as a polynomial in $c(n+1)$ and $c(i)$ for $1 \le i < n$.

Case 2. $c_g^-(2n+1)$ for $2n+1 \not\equiv \pm 1 \bmod p$. By looking at the coefficient of r^2q^{2n} we can express $c_g^-(2n+1)$ in terms of $c_g^-(4n)$ and known quantities. But $c_g^-(4n)$ can be evaluated by looking at the coefficient of r^4q^n and using the fact that $c_{4,n} = c_{3,n+1} = c_{2,n+2}$ (as $2n+1 \not\equiv \pm 1 \bmod p$).

The recursion formulas given by cases 1 or 2 can be written out explicitly and are given in [B, (9.1)]. In that paper they are sufficient to determine all coefficients because the analogues of the relations in proposition 3.4 hold whenever $mn \ne 0$.

The remaining cases of the proof are unusually tedious so we only sketch them briefly.

Case 3. $c_g^-(2n)$ for $2n \equiv 0 \bmod p$, n odd. We can express this in terms of $c_{2,n} = c_{3,n-1}$, which we can evaluate by calculating $c_g^-(3(n-1)) = c_g^-(2(3(n-1)/2))$ using case 1.

For the rest of the proof it is convenient to evaluate the cluster of elements $c_g^-(2pn-1)$, $c_g^-(2pn+1)$, and $c_g^-(4pn)$, simultaneously. (These are the only cases left to do.)

Case 4. $2np \equiv 2 \bmod 3$. We evaluate $c_g^-(2np+1)$ in terms of $c_{3,(2np+1)/3}$. We evaluate $c_g^-(2np-1)$ by expression $c_g^-(2np-1)c(2)$ in terms of $c_{2,2np+1}$ and evaluating the latter by comparing it with $c_{3,2(2np+1)/3}$. We evaluate $c_g^-(4np)$ by expressing $c_{3,2np+2}$ in terms of $c_g^-(2)c_{2,2np}$ and evaluating the latter by comparing it with $c_{2,3np+3}$.

Case 5. $2np \equiv 1 \bmod 3$. We evaluate $c_g^-(2np-1)$ in terms of $c_{3,(2np-1)/3}$. We evaluate $c_{1,2np+5}$ by using $c_{3,(2np+5)/3}$ and then evaluate in turn $c_{2,2np+4} = c_{1,2np+5}$, $c_g^-(1)c_{1,2np+3}$ (using the coefficient of r^2q^{2np+4}), $c_{2,2np+3} = c_{1,2np+3}$, $c_g^-(1)c_{1,2np+1}$ which gives us the value of $c_{1,2np+1}$. We evaluate $c_{2,2np}$ as in case 4.

Case 6. $2np \equiv 0 \bmod 3$. By looking at $c_{2,2np+2}$ and $c_{2,2np+3}$ we can find expressions for the linear combinations $c_g^-(1)c_g^-(2np+1) + c_g^-(3)c_g^-(2np-1)$ and $c_g^-(2)c_g^-(2np+1) + c_g^-(4)c_g^-(2np-1)$. If $p \ne 59$ or 71 then these two linear equations are independent so we can solve for $c_g^-(2np+1)$ and $c_g^-(2np-1)$, and then find $c_{2,2np}$ as in case 4. If $p = 59$ or 71 then an even more tedious argument using $c_{2,2np+2}$, $c_{3,2np+2}$, and $c_{3,2np+3}$ produces 3 independent linear relations between $c_g^-(2pn-1)$, $c_g^-(2pn+1)$, and $c_g^-(4pn)$, so we can solve for them.

This proves lemma 4.7.

We can now complete the proof of theorem 4.1. If we look at the possible solutions for the coefficients $c_g^-(n)$ given by proposition 4.6, we see that all the solutions except those congruent mod 3^{29} to the Hauptmoduls of the elements $g \in M$ are ruled out by the inequalities of lemma 4.2. As $3^{29} > c_1^+(i) > c_g^-(i) \ge 0$ for $i \le 5$ we see that $c_g^+(i) = c_g^-(i)$ for $i \le 5$. The coefficients $c_g^+(n)$ also satisfy the identities of proposition 4.3 satisfied by the $c_g^-(n)$'s; for example, this follows easily from the identity

$$\sum_{n\in\mathbf{Z}} c_g^+(n)r^n - \sum_{n\in\mathbf{Z}} c_g^+(n)q^n = r^{-1} \prod_{m>0,n\in\mathbf{Z}} (1-r^mq^n)^{c_g^+(mn)}(1-r^{pm}q^{pn})^{c_g^+(pmn)}$$

[B, page 434]. Therefore $c_g^-(n) = c_g^+(n)$ for all n by lemma 4.7 because they both have the same values for $n \le 5$ and both satisfy the same recursion relations given in proposition 3.4. This proves theorem 4.1.

5. Calculation of some cohomology groups.

In this section we calculate the Tate cohomology groups of some spaces that appear in the construction of the monster vertex algebra. In the next section we will use these calculations to find the Tate cohomology $\hat{H}^*(g, V[1/2])$ for elements g of types $3B$, $5B$, $7B$ and $13B$.

We will write $g_0, g_1, \dots, g_{p-1}$ for the elements $1, g, \dots, g^{p-1}$ of the group ring $\mathbf{Z}_p[G]$ to avoid confusing the group ring multiplication with the multiplication in other rings containing it.

Suppose that A is a free $\mathbf{Z}_p$-module. We write $h(A)$ for the $\mathbf{Z}$-graded $\mathbf{Z}_p$-algebra generated by elements $h_n(a)$ of degree n for $a \in A$, $n \geq 0$ with the relations

1. $h_n(\lambda a) = \lambda^n h_n(a)$ if $\lambda \in \mathbf{Z}_p$.
2. $h_0(a) = 1$, $h_n(a) = 0$ if $n < 0$.
3. $h_n(a+b) = \sum_{i \in \mathbf{Z}} h_i(a) h_{n-i}(b)$.

If A has a basis $a_1, \dots, a_m$ then $h(A)$ is the polynomial algebra generated by elements $h_n(a_i)$ for $n > 0$, $1 \leq i \leq m$. An action of G on A induces an action of G on $h(A)$ in the obvious way.

Lemma 5.1. *The cohomology ring $\hat{H}^*(g, h(\mathbf{Z}_p[G]))$ is the polynomial algebra over $\mathbf{F}_p$ generated by the ordinary elements $h_n(g_0) + h_n(g_1) + \cdots + h_n(g_{p-1})$ for $n > 0$.*

Proof. This follows from the fact that $h(\mathbf{Z}_p[G])$ is the tensor product of the rings A_n, where A_n is the polynomial algebra generated by the elements $h_n(g_i)$ for $0 \leq i < p$. We see that A_n is isomorphic to the symmetric algebra $S^*(\mathbf{Z}_p[G])$ whose cohomology is given by lemma 2.8, and we then apply corollary 2.2 to finish the proof. (Notice that the space spanned by the elements $h_n(g_i)$ for $0 \leq i < p$ is not $h_n(\mathbf{Z}_p[G])$; in fact, the latter is not even a $\mathbf{Z}_p$-module.) This proves lemma 5.1.

Proposition 5.2. *The cohomology ring $\hat{H}^*(g, h(I))$ is isomorphic to the free exterior algebra with one super generator in each positive degree not divisible by p. The automorphism -1 of I multiplies each generator by -1 and hence acts as -1 on $\hat{H}^1(g, h(I))$ and as 1 on $\hat{H}^0(g, h(I))$.*

Proof. The algebra $h(I)$ is the quotient of $h(\mathbf{Z}_p[G])$ by the ideal generated by the elements $h_n(N_G)$. We let R_m be the quotient of $h(\mathbf{Z}_p[G])$ by the elements $h_n(N_G)$ for $n \leq m$, and we will prove by induction on m that $\hat{H}^*(R_m)$ is the tensor product of the free polynomial algebra generated by the ordinary elements $h_n(g_0) + h_n(g_1) + \cdots + h_n(g_{p-1})$ (of degree n) for $n > m/p$ and the free exterior algebra generated by super elements of each degree n with $0 < n \leq m$, $(p, n) = 1$ (of degree n).

To do this we look at the exact sequence

$$0 \to R_{m-1} \to R_{m-1} \to R_m \to 0$$

where the map from R_{m-1} to R_{m-1} is given by multiplication by $h_m(N_G)$, and work out the cohomology of R_m using the cohomology exact sequence of this short exact sequence. We have to consider the cases $p|m$ and $(p, m) = 1$ separately.

If $(p, m) = 1$ then multiplication by $h_m(N_G)$ induces the zero map on cohomology, so that $\hat{H}^*(g, R_m)$ is the sum of $\hat{H}^*(g, R_{m-1})$ and $\hat{H}^*(g, R_{m-1})$ with the degree shifted by m and with the ordinary and super parts exchanged. In other words, $\hat{H}^*(g, R_m)$ is the tensor product of $\hat{H}^*(g, R_{m-1})$ with an exterior algebra generated by one super element of degree m. This proves the induction step when $(p, m) = 1$.

If $p|m$ then multiplication by $h_m(N_G)$ induces an injective map on cohomology, so that $\hat{H}^*(g, R_m)$ is the quotient of $\hat{H}^*(g, R_{m-1})$ by the image of $h_m(N_G)$, in other words it is just $\hat{H}^*(g, R_{m-1})$ with a polynomial generator of degree m killed off.

Hence we see that to go from $\hat{H}^*(h(R_{m-1}))$ to $\hat{H}^*(h(R_m))$ we either kill off a polynomial generator or add an exterior algebra generator. This proves the statement about $\hat{H}^*(h(R_m))$ for all m by induction, and hence proves proposition 5.2.

For completeness we remark that the cohomology ring $\hat{H}^*(g, h(\mathbf{Z}_p))$ is a polynomial algebra over $\mathbf{F}_p$ with one ordinary generator of each positive degree.

For any $\mathbf{Z}_p$-module A we define ω to be the automorphism of $h(A)$ taking $h_n(a)$ to $(-1)^n h_n(-a)$. We define $h_\omega(A)$ to be the largest quotient ring of $h(A)$ on which the automorphism ω acts trivially. (Notice that this is smaller than the largest quotient module on which ω acts trivially.)

Lemma 5.3. *The cohomology ring $\hat{H}^*(g, h_\omega(\mathbf{Z}_p[G]))$ is the polynomial ring on ordinary generators of degrees $(2n+1)p$ for $n \geq 0$.*

Proof. The algebra $h_\omega(\mathbf{Z}_p[G])$ is the polynomial algebra generated by the elements $h_{2n+1}(g_i)$ for $0 \leq i < p$, $n \geq 0$, and lemma 5.3 follows from this.

Lemma 5.4. *The cohomology ring $\hat{H}^*(g, h_\omega(I))$ is the exterior algebra on generators of degrees $(2n+1)$ for $n \geq 0$, $(n,p) = 1$. In particular it is generated (as a ring) by super elements of odd degrees, so that $\hat{H}^0(g, h_\omega(I))$ vanishes in odd degrees and $\hat{H}^1(g, h_\omega(I))$ vanishes in even degrees.*

This follows from lemma 5.3 in the same way that lemma 5.2 follows from 5.1.

Remark. The ring $h(\mathbf{Z}_p)$ can be identified with the ring of symmetric functions over $\mathbf{Z}_p$ if we identify $h_n(1)$ with the n'th complete symmetric function. In this case the automorphism ω taking $h_n(1)$ to $(-1)^n h_n(-1)$ is just the automorphism ω of [Mac, p. 14] taking the complete symmetric functions to the elementary symmetric functions.

Remark. More generally we can form a quotient ring associated to any automorphism ω of A of order n prime to p such that $h_\omega(A)$ is the quotient ring associated to the automorphism $\omega = -1$ of A. To do this we let R_p be the unramified extension of $\mathbf{Z}_p$ generated by a primitive n'th root of unity ζ. We then form the largest quotient ring of $h(A) \otimes R_p$ whose elements are fixed by the automorphism taking $h_n(a)$ to $\zeta^n h_n(\omega(a))$. For example this space appears when we try to construct the monster vertex algebra from a fixed point free automorphism ω of the Leech lattice Λ.

The extraspecial group 2^{1+24} occurring in the centralizer of an element of type $2B$ in M has a unique 2^{12}-dimensional irreducible module A over $\mathbf{Z}[1/2]$. Any element of odd order in $\mathrm{Aut}(\Lambda)$ has an induced action on 2^{1+24} and on A.

Lemma 5.5. *If g is an element of order 3, 5, 7, or 13 acting on the Leech lattice Λ with no nonzero fixed vectors then as a $\mathbf{Z}_p[G]$-module, $A \otimes \mathbf{Z}_p$ is the sum of $\mathbf{Z}_p$ and some copies of $\mathbf{Z}_p[G]$.*

Proof. We let B be the 2^{24}-dimensional 2^{1+24}-module $A \otimes A$. This is isomorphic to the group ring of 2^{24}. The element g acts fixed point freely on the group $2^{24} = \Lambda/2\Lambda$ because it acts fixed point freely on Λ and has odd order, so as a module over G, $B \otimes \mathbf{Z}_p$ is the sum of $\mathbf{Z}_p$ and some copies of $\mathbf{Z}_p[G]$, and the Tate cohomology groups are therefore $\hat{H}^0(g, B \otimes \mathbf{Z}_p) = \mathbf{Z}/p\mathbf{Z}$, $\hat{H}^1(g, B \otimes \mathbf{Z}_p) = 0$. As $\hat{H}^*(g, B \otimes \mathbf{Z}_p) = \hat{H}^*(g, A \otimes \mathbf{Z}_p) \otimes \hat{H}^*(g, A \otimes \mathbf{Z}_p)$ we see that $\hat{H}^*(g, A \otimes \mathbf{Z}_p)$ is either an ordinary or a super vector space over $\mathbf{F}_p$ of dimension 1.

Therefore as a G module, $A \otimes \mathbf{Z}_p$ is the sum of some copies of $\mathbf{Z}_p[G]$ and either $\mathbf{Z}_p$ or I. But $\dim(A) = 2^{12} \equiv 1 \bmod p$ as $(p-1)|12$ which rules out the second possibility. Therefore $\hat{H}^0(g, A \otimes \mathbf{Z}_p)) = \mathbf{Z}/p\mathbf{Z}$, $\hat{H}^1(g, A \otimes \mathbf{Z}_p) = 0$. This proves lemma 5.5.

A similar argument can be used to calculate $\hat{H}^*(g, A)$ for any element g of odd prime order p in $\mathrm{Aut}(\Lambda)$. If $2n$ is the dimension of the subspace of Λ fixed by g then n is an integer with $2^n \equiv \pm 1 \bmod p$. Exactly one of the two cohomology groups $\hat{H}^*(g, A)$ vanishes and the other has dimension 2^n over $\mathbf{F}_p$. The group $\hat{H}^1(g, A)$ vanishes if $2^n \equiv 1 \bmod p$, and the group $\hat{H}^0(g, A)$ vanishes if $2^n \equiv -1 \bmod p$. There are some cases where it is $\hat{H}^0(g, A)$ and not $\hat{H}^1(g, A)$ that vanishes; for example, if g is the element of order 5 with trace -1 on Λ then $\hat{H}^0(g, A) = 0$ and $\hat{H}^1(g, A) = (\mathbf{Z}/5\mathbf{Z})^4$.

6. Elements of type $3B$, $5B$, $7B$, and $13B$.

We use the result of the previous section to calculate the cohomology groups $\hat{H}^i(g, V[1/2])$ when $g \in M$ has type $3B$, $5B$, $7B$, or $13B$.

Theorem 6.1. *If $g \in M$ has type $3B$, $5B$, $7B$, or $13B$ and σ is the element of order 2 generating the center of $C_M(g)/O_p(C_M(g))$ then $O_p(C_M(g))$ acts trivially on ${}^gV$ and σ fixes the ordinary part of ${}^gV$ and acts as -1 on the super part.*

Proof. We prove this by calculating the action of σ on $\hat{H}^*(g, V[1/2])$ explicitly. We recall from [B-R] that the $\mathbf{Z}[1/2]$-form $V[1/2]$ of Frenkel, Lepowsky and Meurman's monster vertex algebra [FLM] is the sum of two pieces V^0 and V^1 which are the $+1$ and -1 eigenspaces of an element σ of type $2B$. (Unfortunately the details for the construction of this $\mathbf{Z}[1/2]$-form have not been published anywhere.) We now show that $\hat{H}^0(g, V^1) = \hat{H}^1(g, V^0) = 0$.

The space V^0 is the subspace of $V_\Lambda[1/2]$ fixed by the automorphism induced by -1 acting on the Leech lattice Λ, where $V_\Lambda[1/2] = V_\Lambda \otimes \mathbf{Z}[1/2]$, and V_Λ is the integral form of the vertex algebra of the Leech lattice, described in [B-R]. As in [B-R], the space V_Λ is the tensor product of a twisted group ring $\mathbf{Z}[\hat{\Lambda}]$ and a space isomorphic to $h(\Lambda)$. The cohomology $\hat{H}^*(g, \mathbf{Z}[\hat{\Lambda}])$ is just $\mathbf{Z}/p\mathbf{Z}$ as in [B-R] because g acts fixed point freely

on Λ. By proposition 5.2, the element -1 of $\mathrm{Aut}(\hat{\Lambda})$ acts as $+1$ on $\hat{H}^0(g, h(\Lambda))$ and as -1 on $\hat{H}^1(g, h(\Lambda))$. Therefore $\hat{H}^1(g, V^0)$ vanishes because it is isomorphic to the subspace of $\hat{H}^1(g, h(\Lambda))$ fixed by $-1 \in \mathrm{Aut}(\Lambda)$.

The space V^1 is the subspace of elements of integral weight of the tensor product $A \otimes h_\omega(\Lambda \otimes \mathbf{Z}_p)$, where A is the 2^{12} dimensional module for 2^{1+24} and $h_\omega(\Lambda \otimes \mathbf{Z}_p)$ has its grading halved and then shifted by 3/2. By lemma 5.5, $\hat{H}^*(g, A) = \mathbf{F}_p$, so $\hat{H}^*(g, A \otimes h_\omega(\Lambda \otimes \mathbf{Z}_p)) = \hat{H}^*(g, h_\omega(\Lambda \otimes \mathbf{Z}_p))$. By lemma 5.4 the cohomology ring $\hat{H}^*(g, h_\omega(\Lambda \otimes \mathbf{Z}_p))$ is generated by super elements of odd degree, so $\hat{H}^1(g, h_\omega(\Lambda \otimes \mathbf{Z}_p))$ vanishes in even degrees and $\hat{H}^1(g, h_\omega(\Lambda \otimes \mathbf{Z}_p))$ vanishes in odd degrees. Therefore $\hat{H}^0(g, A \otimes h_\omega(\Lambda \otimes \mathbf{Z}_p))$ vanishes in integral degrees, so $\hat{H}^0(g, V^1) = 0$.

This shows that the element σ acts as 1 on $\hat{H}^0(g, V[1/2]) = \hat{H}^0(g, V_\Lambda) = \hat{H}^0(g, h(\Lambda \otimes \mathbf{Z}_p))$ and as -1 on $\hat{H}^1(g, V[1/2]) = \hat{H}^1(g, A \otimes h_\omega(\Lambda \otimes \mathbf{Z}_p))$.

Finally we show that $O_p(C_M(g))$ acts trivially on ${}^gV$. The element σ acts trivially on the center $\langle g \rangle$ of the extraspecial group $p^{1+24/(p-1)}$ and acts as -1 on the quotient $p^{24/(p-1)}$. The element g acts trivially on ${}^gV$ so we get an action of $p^{24/(p-1)}$ on ${}^gV$. As σ acts as 1 on the even part of ${}^gV$ and as -1 on the odd part, it commutes with all automorphisms of ${}^gV$, so the group $p^{24/(p-1)}$ must act trivially on ${}^gV$. This proves theorem 6.1.

The group $C_M(g)$ has structure $3^{1+12}.2.Suz$, $5^{1+6}.2.HJ$, $7^{1+4}.2.A_7$, or $13^{1+2}.2.A_4$, so we get modular vertex superalgebras acted on faithfully by the groups $2.Suz$, $2.HJ$, $2.A_7$, and $2.A_4$ whose modular characters are given by Hauptmoduls. We can calculate the characters of both the ordinary and super parts of these superalgebras by using the fact that the characters of their sum and difference are both given in terms of known Hauptmoduls.

It seems likely that if g has type $2B$ then the vertex superalgebra acted on by $2^{24}.Co_1$ constructed in [B-R] is acted on trivially by the 2^{24} subgroup, but this has not yet been proved. The proof given above for the corresponding fact for elements of odd prime order fails for order 2 because an automorphism mapping each element of 2^{24} to its inverse is still the identity automorphism.

As an example, we work out the dimensions of the ordinary and super parts of ${}^gV$ in the case when g has type $3B$. The function $\sum_n \mathrm{Tr}(1|{}^gV)$ is equal to the Hauptmodul of the element g of type $3B$ in the monster, which is

$$q^{-1} + 54q - 76q^2 - 243q^3 + 1188q^4 - 1384q^5 - \cdots,$$

and the function $\sum_n \mathrm{Tr}(\sigma|{}^gV)$ is equal to the Hauptmodul of the element σg of type $6B$ in the monster, which is

$$q^{-1} + 78q + 364q^2 + 1365q^3 + 4380q^4 + 12520q^5 + \cdots$$

so we find that the dimension of the ordinary part of ${}^gV_n$ is the coefficient of q^n of

$$q^{-1} + 66q + 144q^2 + 561q^3 + 2784q^4 + 5568q^5 + \cdots$$

and the dimension of the super part of ${}^gV_n$ is the coefficient of q^n of

$$12q + 220q^2 + 804q^3 + 1596q^4 + 6952q^5 \cdots.$$

These series can be written as sums of dimensions of ordinary representations of the groups $3.Suz$ and $6.Suz$ as

$$q^{-1}+66q+(1+143)q^2+(429+66+66)q^3+(66+66+78+429+2145)q^4+(4\times1+4\times143+2\times780+3432)q^5+\cdots$$

and

$$12q + 220q^2 + (780 + 12 + 12)q^3 + (780 + 780 + 12 + 12 + 12)q^4 + (4 \times 220 + 2 \times 572 + 4928)q^5 + \cdots.$$

(Some of the decompositions are ambiguous; for example $1 + 143$ could be $66 + 78$. We have chosen $1 + 143$ instead of $66 + 78$ for a reason that will appear in a moment.) At first sight this seems to rule out a vertex algebra structure on any lift to characteristic 0, as the homogeneous components V_n, $V_{n+1}, \ldots$ have to increase. However we can get around this by rescaling the grading by a factor of 1/3. The data above seem consistent with the existence of a $\frac{1}{3}\mathbf{Z}$-graded vertex superalgebra $W = \sum_{3n \in \mathbf{Z}} W_n$ such that the mod

3 reduction of W_n is V_{3n}. The vertex superalgebra W is probably acted on by $6.Suz$, with the element of order 3 in the center acting as ω^{3n} on W_n.

There is a second modular vertex super algebra acted on by each of the groups $2.Suz$, $2.HJ$, $2.A_7$, and $2.A_4$, which can be constructed as $\hat{H}^*(g, V_\Lambda)$ where g is a fixed point free automorphism of order 3, 5, 7, or 13 of the Leech lattice Λ. The ordinary part of this superalgebra is the same as the ordinary part of the superalgebra constructed from the monster vertex algebra, but its super part is different; for example, its degree 0 piece does not vanish.

7. Open problems and conjectures.

For more open problems in this area see section 7 of [B-R].

At the end of section 6 we saw that the vertex superalgebra of an element of type $3B$ looks as if it might be the reduction mod 3 of a superalgebra in characteristic 0. More generally something similar seems to happen for some other elements of the monster, as described in the following conjecture.

Conjecture. *For each element $g \in M$ of order n there is a $\frac{1}{n}\mathbf{Z}$-graded super module ${}^g\hat{V} = \sum_{i\in\frac{1}{n}\mathbf{Z}} V_i$ over the ring $\mathbf{Z}[e^{2\pi i/n}]$ with the following properties.*

1. *${}^g\hat{V}$ is often a vertex superalgebra with a conformal vector and a self dual invariant symmetric bilinear form. (For example, it seems to be a vertex algebra when g is of type $1A$, $2B$, $3B$, $5B$, $7B$, $13B$, or $3C$, but possibly not if g has type $2A$ or $3A$.)*
2. *${}^1\hat{V}$ is a self dual integral form of the monster vertex algebra (probably the one conjecturally described in lemma 7.1 of [B-R]).*
3. *If g has prime order then the reduction mod p of ${}^g\hat{V}$ is the modular vertex superalgebra ${}^gV$ associated to g (at least when its grading is multiplied by n.)*
4. *If all coefficients of the Hauptmodul of g are nonnegative then the super part of ${}^g\hat{V}$ vanishes. In this case ${}^g\hat{V}$ should be the same as the "twisted sector" of g that appears in Norton's generalized moonshine conjectures [N]. (But notice that if the Hauptmodul of g has negative coefficients then ${}^g\hat{V}$ is certainly not the twisted sector of g; for example, ${}^g\hat{V}$ is a super vector space and has a nontrivial piece of degree 0.)*
5. *${}^g\hat{V}$ is acted on by a central extension $(\mathbf{Z}/n\mathbf{Z}).C_M(g)$ of $C_M(g)$ by $\mathbf{Z}/n\mathbf{Z}$. This central extension contains an element of order 1 or 2 acting as -1 on the super part of ${}^g\hat{V}$ and as 1 on the ordinary part.*
6. *If g and h are commuting elements of M of coprime orders and $\hat{g}$ is the lift of g to the central extension of $C_M(h)$ with the same order as g, then*

$$\mathrm{Tr}(\hat{g}|^h\hat{V}) = \mathrm{Tr}(gh|^1\hat{V}).$$

   (This identity is often false if the orders of g and h are not coprime.)
7. *If g and h commute then $\hat{H}^*(\hat{g},{}^h\hat{V})$ is (essentially) the reduction mod $|g|$ of ${}^{gh}V$. (Actually this is not quite correct as stated, because the reduction mod $|g|$ will be defined over some nilpotent extension of $\mathbf{Z}/|g|\mathbf{Z}$, and we should then take the tensor product over this ring with $\mathbf{Z}/|g|\mathbf{Z}$.)*

These conjectures are similar to Norton's generalized moonshine conjectures [N] because both associate some graded space with each element g of the monster. However they treat the problem of negative coefficients in the Hauptmodul $T_g(\tau)$ in a different way: Norton's conjectures use the fact that $T_g(-1/\tau)$ has nonnegative coefficients, which are supposed to be the dimensions of some vector spaces, while the conjectures above allow super vector spaces which may have negative super dimension. Perhaps both Norton's conjectures and the conjectures above are special cases of some conjecture which associates a $\mathbf{Z}/n\mathbf{Z}\times\mathbf{Z}$-graded super space acted on by a central extension $\mathbf{Z}/n\mathbf{Z}.C_M(g)$ to each element $g \in M$ of order n.

We finish by listing a few other questions.

1. Is the group $C_M(g)/O_p(C_M(g))$ the full automorphism group of the modular vertex algebra ${}^gV$? If g has type $3C$ it seems possible that it is not, and the full automorphism group may be $E_8(\mathbf{F}_3)$. Notice that if g has type $2B$ then it is not yet known that $O_2(C_M(g))$ acts trivially on ${}^gV$.
2. Replace the revolting argument of section 4 that involves computer calculations with a more conceptual proof. More generally, suppose we have coefficients $c(n)$ of some function such that the coefficient of $r^m q^n$ in

$$r^{-1}\prod_{m>0,n\in\mathbf{Z}}(1-r^mq^n)^{c(mn)}$$

vanishes whenever $(mn, N) = 1$ for some fixed integer N. What does this imply about the coefficients $c(n)$? (There are many solutions to this given by taking the $c(n)$'s to be the coefficients of some Hauptmodul of level N.)

3. What happens when we apply the argument of section 3 to some automorphism of prime order of the fake monster Lie algebra? Do the infinite products we get represent interesting functions?
4. In [B section 10] it was suggested that there should be Lie superalgebras associated with elements g of the monster, but it was not clear what the dimensions of homogeneous components of these superalgebras should be in the cases when the coefficients of the Hauptmodul of g were neither all positive nor alternating in sign. The results of section 6 suggest what to do in this case: the sum of the dimensions of the ordinary and super parts should also be given by the coefficients of another Hauptmodul.

References.

[BBCO] C. Batut, D. Bernardi, H. Cohen, M. Olivier, "User's guide to PARI-GP". This guide and the PARI programs can be obtained by anonymous ftp from hensel.mathp7.jussieu.fr in the directory /dist/pari.

[B] R. E. Borcherds, Monstrous moonshine and monstrous Lie superalgebras, Invent. Math. 109, 405-444 (1992).

[B-R] R. E. Borcherds, A. J. E. Ryba, Modular Moonshine II, Duke Math Journal Vol. 83 No. 2, 435-459, 1996.

[CG] C. J. Cummins, T. Gannon, Modular equations and the genus zero property of moonshine functions, Invent. Math. 129, 413-443 1997.

[DM] C. Dong and G. Mason, On the construction of the moonshine module as a $\mathbf{Z}_p$-orbifold, Mathematical aspects of conformal and topological field theories and quantum groups (South Hadley, MA, 1992), 37–52, Contemp. Math., 175, Amer. Math. Soc., Providence, RI, 1994.

[FLM] I. B. Frenkel, J. Lepowsky, A. Meurman, Vertex operator algebras and the monster, Academic press 1988.

[Mac] I. G. Macdonald, "Symmetric functions and Hall polynomials", Oxford University press, 1979.

[Ma] Y. Martin, On modular invariance of completely replicable functions. Moonshine, the Monster, and related topics (South Hadley, MA, 1994), 263–286, Contemp. Math., 193, Amer. Math. Soc., Providence, RI, 1996.

[M] P. Montague, Third and Higher Order NFPA Twisted Constructions of Conformal Field Theories from Lattices, Nuclear Phys. B 441 (1995), no. 1-2, 337–382.

[N] S. P. Norton, Generalized Moonshine, Proc. Symp. Pure Math. 47 (1987) p. 208-209.

[R] A. J. E. Ryba, Modular Moonshine?, In "Moonshine, the Monster, and related topics", edited by Chongying Dong and Geoffrey Mason. Contemporary Mathematics, 193. American Mathematical Society, Providence, RI, 1996. 307-336.

The fake monster formal group. *27 May, 31 Dec 1998*

Richard E. Borcherds, *
D.P.M.M.S., 16 Mill Lane, Cambridge, CB2 1SB, England.
e-mail: reb@dpmms.cam.ac.uk
home page: www.dpmms.cam.ac.uk/~reb

arXiv:math/9805123v2 [math.QA] 22 Jan 1999

Contents.

1. Introduction.
Notation and terminology.
2. Some theorems about smooth Hopf algebras.
3. Liftings of Lie algebra elements.
4. The smooth fake monster Hopf algebra.
5. A smooth Hopf algebra for the Virasoro algebra.
6. The no-ghost theorem over **Z**.
7. An application to modular moonshine.
8. Open problems.

1. Introduction.

The main result of this paper is the construction of "good" integral forms for the universal enveloping algebras of the fake monster Lie algebra and the Virasoro algebra. As an application we construct formal group laws over the integers for these Lie algebras. We also prove a form of the no-ghost theorem over the integers, and use this to verify an assumption used in the proof of the modular moonshine conjectures.

Over the integers the universal enveloping algebra of a Lie algebra is not very well behaved, and it is necessary to use a better integral form of the universal enveloping algebra over the rational numbers. The correct notion of "good" integral form was found by Kostant [K]. He found that the good integral forms are the ones with a structural base (as defined in 2.3), and showed that finite dimensional semisimple Lie algebras have a structural base. The existence of a structural base implies that the dual algebra of the underlying coalgebra is a ring of formal power series. As this ring can be thought of as a sort of "coordinate ring" of some sort of formal group, this condition can be though of as saying that the formal group is smooth and connected, and that the formal group comes from a "formal group law".

The main point of this paper is to find such integral forms for the universal enveloping algebras of certain infinite dimensional Lie algebras. We recall that the universal enveloping algebra of a Lie algebra has a natural structure of a cocommutative Hopf algebra. So we need some theorems to tell us when a Hopf algebra has a structural base. In section 2 we prove that a Hopf algebra H has a structural base provided that there are "sufficiently many" group-like elements $1 + a_1x + a_2x^2 + \cdots \in H[[x]]$, and provided a few other minor conditions are satisfied. Such a group-like element should be thought of as an infinitesimal curve in the formal group, so roughly speaking this condition means that a set of generators of the Lie algebra of the formal group can be lifted to formal curves in the formal group.

* Supported by a Royal Society professorship and by an NSF grant.

As an example of the theorems in section 2, we find that the Lie bracket $[a_1, b_1]$ of two primitive liftable elements is also liftable. At first sight this looks as if it should be easy to prove, as all we need to do is write down an explicit lifting, with coefficients that are some universal (non-commutative) polynomials in the coefficients of liftings of a_1 and b_1. However this seems to be very hard to do explicitly (partly because such a lifting is far from unique, which paradoxically makes it harder to find one). Instead, we use a far more roundabout argument, where we first need to prove a theorem (theorem 2.12) saying that all primitive elements of certain Hopf algebras are liftable.

We also give an incidental application of these theorems to certain Hopf algebras F_n considered by Dieudonné. He showed that these Hopf algebras have a structural basis when reduced mod p for any prime p; we show that the Hopf algebras F_n already have a structural basis over the integers.

In section 2 we have reduced the problem of finding a structural basis of a Hopf algebra H to the problem of finding enough group-like elements in $H[[x]]$. For the universal enveloping algebras of finite dimensional simple Lie algebras, or more generally for Kac-Moody algebras, this is easy to do, because we just take the formal 1-parameter subgroup

$$\sum_{n\geq 0} x^n Ad(e)^n/n!$$

for (locally nilpotent) elements e of the Lie algebra corresponding to real roots, and define the integral form of the Hopf algebra to be the one generated by the coefficients of these liftings. This gives the usual Kostant integral form for the universal enveloping algebra of finite dimensional semisimple Lie algebras. For the Lie algebras in this paper this does not work because there are not enough locally nilpotent elements. (We can try to use the elements $e^n/n!$ for non nilpotent elements, but there seems no obvious way in which this gives a good integral form.) Fortunately we do not need to lift the generators of the Lie algebra to formal one parameter groups, and it is sufficient to lift them to formal curves, which is easier to do.

In section 3 we construct liftings of some elements of the vertex algebra of a lattice to formal curves. More precisely, we show that we can lift any element of a certain Lie algebra provided that it lies in a root space of a root of norm 2 or 0. We do this by explicitly writing down a group-like lifting, and checking by brute force that its coefficients are integral. (Unfortunately there seems no obvious way to extend this brute force approach to root spaces of negative norm roots.)

In section 4 we use the liftings of section 3 to construct a smooth integral form of the universal enveloping algebra of the fake monster Lie algebra, which is constructed from the lattice $II_{25,1}$. The main point about this Lie algebra is that it is generated by the root spaces of norm 2 and norm 0 roots, so in section 3 we have constructed enough liftings to apply the theorems in section 2. (This is a very special property of $II_{25,1}$; it is the only known indefinite lattice whose Lie algebra is generated by the root spaces of norm 2 and norm 0 roots. In other words the theory in sections 2 and 3 has been developed mainly for this one example!) We can summarize the main results about the fake monster Lie algebra proved in this paper as follows.

Theorem 4.1. *There is a $II_{25,1}$-graded Hopf algebra $U^+(\mathbf{m})$ over $\mathbf{Z}$ with the following properties.*

1. *$U^+(\mathbf{m})$ has a structural basis over $\mathbf{Z}$.*
2. *The primitive elements of $U^+(\mathbf{m})$ are an integral form of the fake monster Lie algebra $\mathbf{m}$.*
3. *For every norm 2 vector of $II_{25,1}$, $U^+(\mathbf{m})$ contains the usual (Kostant) integral form of the universal enveloping algebra of the corresponding $sl_2(\mathbf{Z})$.*

In section 5 we construct a smooth integral form for the universal enveloping algebra of the Virasoro algebra, by applying the theorems in section 2 to a set of explicit liftings of the basis elements of the Virasoro algebra. In other words we construct a formal group law for the Virasoro algebra over the integers.

The no-ghost theorem in string theory states that a certain real vector space of states is positive definite (so it contains no negative norm vectors, which are sometimes called ghosts and which would prevent the space from being a Hilbert space). This real vector space has a natural integral form which can be made into a positive definite lattice using the inner product, and we can ask for an integral form of the no-ghost theorem, which should say something about the structure of this lattice. In section 6 we use the smooth formal group of the Virasoro algebra to prove an integral form of the no-ghost theorem, at least in the case of vertex algebras constructed from 26 dimensional lattices. More precisely we prove bounds on the determinant of this lattice, which quite often imply that the lattice is self dual.

The proof [B98] of Ryba's modular moonshine conjectures for large primes used an unproved technical assumption about the monster Lie algebra. In section 7 we use the integral no-ghost theorem to prove this assumption, thus completing the proof of the modular moonshine conjectures for odd primes. (The proof in [B-R] for the prime 2 still relies on another unproved technical assumption.)

Notation and terminology.

α	An element of a lattice.
β	An element of a lattice.
γ	A norm 0 element of a lattice.
Γ	An indeterminate.
c	An element generating the center of the Virasoro algebra.
Der	The derivations of a commutative ring.
e^α	An element of the group ring of a lattice L, usually regarded as an element of the vertex algebra of L.
ϵ_i	The function from I to $\mathbf{Z}$ taking value 1 on $i \in I$ and 0 elsewhere.
$\mathbf{F}_p$	A finite field of order p.
$F(\lambda)$	If $\lambda = 1^{i_1}2^{i_2}\cdots$ is a partition, then $F(\lambda) = i_1!i_2!\cdots$.
γ	A norm 0 vector of a lattice L.
H	A cocommutative Hopf algebra.
I	An index set, or a finite sequence of integers.
$II_{m,n}$	An even self dual lattice of dimension $m+n$ and signature $m-n$.
J	A finite sequence of integers.

$l(I)$ The length of the sequence I.
$l(\lambda)$ If $\lambda = 1^{i_1}2^{i_2}\cdots$ is a partition, then $l(\lambda) = i_1 + i_2 + \cdots$.
λ A partition.
K A field, sometimes the quotient field of R.
L A lattice, usually even and self dual.
L_i A basis element for the Virasoro or Witt algebra.
$\mathbf{m}$ An integral form of the fake monster Lie algebra (or the monster Lie algebra in section 7).
N An integer.
$P(\lambda)$ If $\lambda = 1^{i_1}2^{i_2}\cdots$ is a partition, then $P(\lambda)$ is the integer $1^{i_1}2^{i_2}\cdots$.
p A prime.
$p(n)$ The number of partitions of n.
$\mathbf{Q}$ The rational numbers.
ρ A Weyl vector.
R A commutative ring.
S The antipode of a Hopf algebra.
$\Sigma(I)$ The sum of the elements of I.
U A universal enveloping algebra.
U^+ An integral form of a universal enveloping algebra with a structural basis.
V A vertex algebra.
V_p The shift (or Verschiebung) of a coalgebra over $\mathbf{F}_p$.
Vir The Virasoro algebra.
$Witt$ The Witt algebra.
x A formal variable.
$\mathbf{Z}$ The integers. $\mathbf{Z}_{\geq 0}$ is the set of non-negative integers.
$\mathbf{Z}_p$ The p-adic integers.
Z_α An element of a structural basis.
z A formal variable.

Bialgebra A module with compatible algebra and coalgebra structures.
Group-like $\Delta(a) = a \otimes a$.
Hopf algebra A bialgebra with antipode.
Liftable See definition 2.2.
Primitive $\Delta(a) = a \otimes 1 + 1 \otimes a$.
Smooth. A Hopf algebra or coalgebra with a structural basis.
Structural basis $\Delta Z_\alpha = \sum_{0 \leq \beta \leq \alpha} Z_\beta \otimes Z_{\alpha-\beta}$

2. Some theorems about Hopf algebras.

In this section we will prove several results about Hopf algebras. The main result is theorem 2.15, which states that under mild conditions a bialgebra generated by the coefficients of group-like liftings has a structural basis. In later sections we will use this to construct a good integral form of the universal enveloping algebra of the fake monster Lie algebra.

We recall that a coalgebra over a commutative ring R is an R-module H with a coassociative coproduct and a counit, a bialgebra is a module with compatible algebra and coalgebra structures, and a Hopf algebra is a bialgebra with an antipode S.

Definition 2.1. *An element a of a coalgebra is called group-like if $\Delta(a) = a \otimes a$, and primitive if $\Delta(a) = a \otimes 1 + 1 \otimes a$.*

Definition 2.2. *We say that an element a_1 of the coalgebra H is liftable to a group-like element, or liftable for short, if we can find elements $a_n \in H$ $(n \geq 0)$ such that $a_0 = 1$ and the element $\sum_{n\geq 0} a_n z^n \in H[[z]]$ is group-like. In other words*

$$\Delta(a_n) = \sum_{0\leq m\leq n} a_m \otimes a_{n-m}$$

for all $n \geq 0$. We say that a_1 is liftable to order N if we can find elements a_n for $0 \leq n \leq N$ (with $a_0 = 1$) satisfying the relations above for $0 \leq n \leq N$. We say that the lifting is graded if H is graded by some abelian group L and $\deg(a_n) = n\alpha$ for some $\alpha \in L$.

We write $\mathbf{Z}_{\geq 0}$ for the nonnegative integers, and $\mathbf{Z}_{\geq 0}^{(I)}$ for the functions from a set I to $\mathbf{Z}_{\geq 0}$ that are zero on all but a finite number of elements of I. The element ϵ_i of $\mathbf{Z}_{\geq 0}^{(I)}$ is defined to be the function that is 1 on $i \in I$ and 0 elsewhere. We define a partial order $\geq$ on $\mathbf{Z}_{\geq 0}^{(I)}$ in the obvious way, by saying $\alpha \geq \beta$ if $\alpha(i) \geq \beta(i)$ for all $i \in I$.

Definition 2.3. *A structural basis for a coalgebra or Hopf algebra H over a ring R is a set of elements Z_α, $\alpha \in \mathbf{Z}_{\geq 0}^{(I)}$ for some set I, such that the elements Z_α form a basis for the free R-module H and such that $\sum_\alpha Z_\alpha x^\alpha$ is group-like, in other words*

$$\Delta(Z_\alpha) = \sum_{0\leq\beta\leq\alpha} Z_\beta \otimes Z_{\alpha-\beta}.$$

Lemma 2.4. *If a coalgebra H has a structural basis Z_α then the R-module of primitive elements of H has a basis consisting of the elements Z_{ϵ_i} for $i \in I$.*

Proof. We have to show that every primitive element of H is a linear combination of the elements Z_{ϵ_i}. Give H the bialgebra structure such that $Z_\alpha Z_\beta = Z_{\alpha+\beta} \prod_{i\in I} \binom{\alpha(i)+\beta(i)}{\beta(i)}$ (as in [A, section 2.5.1] when R is a field). The dual algebra H^* is a ring of formal power series and H acts on H^* as differential operators. Suppose D is any primitive element of H. By subtracting multiples of Z_{ϵ_i} from D we can assume that D acts trivially on the elements x_i dual to Z_{ϵ_i}. But then the fact that D is a derivation implies that D acts trivially on any polynomial in the x_i's. As there are no elements of H orthogonal to all polynomials in the x_i's this proves that D must be 0. This proves lemma 2.4.

Lemma 2.5. *The liftable elements of a bialgebra over R form an R-submodule.*

Proof. If $a(x)$ and $b(x)$ are lifts of a_1 and b_1 then $a(x)b(x)$ is a lift of $a_1 + b_1$, and if $r \in R$ then $a(rx)$ is a lift of ra_1. This proves lemma 2.5.

Lemma 2.6. *If H is a bialgebra with a structural basis then all primitive elements of H are liftable.*

Proof. By lemma 2.4 the R-module of primitive elements has a base of elements Z_{ϵ_i}. Each of these basis elements can be lifted by $\sum_{n\geq 0} Z_{n\epsilon_i}x^n$. Lemma 2.6 now follows from lemma 2.5.

Lemma 2.7. *Suppose H is a coalgebra with a structural basis over an integral domain R. Then any nonzero coideal J of H contains a nonzero primitive element of H.*

Proof. We first assume that R is a field. A coalgebra over a field with a structural basis is irreducible and pointed (by [A, section 2.5]), so lemma 2.7 follows immediately from [A, corollary 2.4.14], which states that any nonzero coideal of a pointed irreducible coalgebra over a field contains a nonzero primitive element.

For the general case, let K be the quotient field of R. If the coideal J is nonzero then $J \otimes_R K$ is a nonzero coideal of $H \otimes_R K$. As lemma 2.7 is is true over the field K this shows that $J \otimes_R K$ contains a primitive element, and multiplying this by a suitable nonzero constant to cancel the denominators gives a nonzero primitive element of J. This proves lemma 2.7.

Lemma 2.8. *Suppose that H is a torsion-free coalgebra over $\mathbf{Z}$ such that for every prime p every primitive element of H/pH is the image of a primitive element of H. Let N be any positive integer. Then every primitive element of H/NH is the image of a primitive element of H.*

Proof. We prove this by induction on the number of prime factors of N, the case $N = 1$ being trivial. Suppose p is any prime factor of N, and suppose that the image of $x \in H$ is primitive in H/NH, so that we have to show that x is congruent to a primitive element mod N. The element x maps to a primitive element in H/pH, so by assumption we have $x = y+pz$ for some primitive element y and some z. But then $\Delta(pz) \equiv pz\otimes 1+1\otimes pz$ mod N as x and y are both primitive mod N. Using the fact that H is torsion-free we can divide through by p to find that $\Delta(z) \equiv z \otimes 1 + 1 \otimes z$ mod N/p. By induction on the number of prime factors of N this shows that $z = t+(N/p)u$ for some primitive element t. Substituting this back in to $x = y + pz$ shows that $x = y + pt + Nu$, so x is congruent to a primitive element $y + pt$ mod N. This proves lemma 2.8.

Lemma 2.9. *If every primitive element of a bialgebra is liftable to all finite orders, then every primitive element is liftable (to infinite order).*

Note that we do not claim that if one particular primitive element of a cocommutative Hopf algebra can be lifted to all finite orders then it can be lifted to infinite order. I do not know whether or not this is true in general.

Proof. Suppose that $1 + a_1x + a_2x^2 + \cdots + a_nx^n$ is an order n lifting of a_1. It is sufficient to show that we can extend this to an order $n + 1$ lifting, because by repeating this an infinite number of times we get a lifting of a_1 of infinite order. We will prove by induction on k that if $k \leq n$ then we can find an order $n+1$ lifting of a_1 which agrees with $a(x)$ to order k. This is true for $k = 1$ by the assumption that a_1 has liftings of arbitrary finite order. Let $b(x) = 1 + b_1x + \cdots + b_{n+1}x^{n+1} \in H[x]/(x^{n+2})$ be an order $n + 1$ lifting

of a_1 with $b_i = a_i$ for $i \le k$. Then $a(x)b(x)^{-1} = 1 + (a_{k+1} - b_{k+1})x^{k+1} + \cdots$ is group-like, so $a_{k+1} - b_{k+1}$ is primitive. (Note that $b(x)^{-1}$ is a well defined order $n+1$ lifting, as we can construct it as $\sum_m (1 - b(x))^m$.) Let $c(x)$ be an order $n+1$ lifting of $a_{k+1} - b_{k+1}$ (which exists as we assumed all primitive elements have liftings of any finite order). Then $b(x)c(x^{k+1})$ is an order $n+1$ lifting of a_1, which agrees with $a(x)$ to order $k+1$. This shows that any order n lifting of a_1 extends to an order $n+1$ lifting. This proves lemma 2.9.

Lemma 2.10. *Suppose that H is a torsion-free bialgebra over $\mathbf{Z}$ and suppose that for every prime p the map from primitive elements of H to primitive elements of H/pH is onto. Then every primitive element of H is liftable.*

Proof. We will show by induction on n that any primitive element a can be lifted to a group-like element in $H[x]/(x^{n+1})$; this is trivial for $n = 0$. We will show by induction on k that if $1 \le k \le n$ then we can find a lifting $\sum_{0 \le i \le n} c_i x^i$ of $n!a$ such that $n!^i | c_i$ for $i \le k$. This is true for $k = 1$, because we can use the order n lifting $\sum_{0 \le i \le n} (n!^i/i!) a^i x^i$. So suppose that $k \ge 2$ and that $n!^i | c_i$ for $i < k$. Then

$$\Delta(c_k) = \sum_{0 \le i \le k} c_i \otimes c_{k-i} \equiv c_k \otimes 1 + 1 \otimes c_k \bmod n!^k.$$

Therefore c_k is primitive in $H/n!^k H$. By lemma 2.8 and the assumption that for every prime p the map from primitive elements of H to primitive elements of H/pH is onto this implies that $c_k = d_1 + n!^k e$ for some primitive d_1 and some e. By induction on n we can find a lifting $d(x) = \sum_{0 \le i < n} d_i x^i \in H[x]/(x^n)$ of d_1, and changing x to x^k we get a group-like element $d(x^k) \in H[x]/(x^{n+1})$. (Here we use the fact that $k \ge 2$ and $n \ge 1$, so that $kn \ge n+1$.) Multiplying $\sum c_i x^i$ by $d(x)^{-1} \bmod x^{n+1}$ we get a group-like element in $H[x]/(x^{n+1})$ whose coefficient of x^i for $i \le k$ is divisible by $n!^k$. This completes the proof by induction on k, and shows that we can assume that $n!^i | c_i$ for all $i \le n$. But then using the fact that H is torsion free we see that $\sum (c_i/n!^i) x^i$ is a lifting of $c_1/n! = a$ in $H[x]/(x^{n+1})$.

We have shown that any primitive element of H can be lifted to a group-like element of any finite order n. By lemma 2.9 this implies that every primitive element of H is liftable. This proves lemma 2.10.

We recall that for any cocommutative coalgebra or Hopf algebra H over $\mathbf{F}_p$ there is a homomorphism V_p from H to itself, called the shift (or Verschiebung), which is dual to the Frobenius map $x \mapsto x^p$ on the dual algebra H^* (see for example [A section 5.4]). The shift V_p commutes with all homomorphisms of coalgebras. If H is a coalgebra over $\mathbf{Z}$ we also write V_p for the shift of H/pH.

Lemma 2.11. *Suppose that the Z_α's for $\alpha \in \mathbf{Z}_{\ge 0}^{(I)}$ are elements of a coalgebra H over $\mathbf{F}_p$ satisfying the relations*

$$\Delta(Z_\alpha) = \sum_{0 \le \beta \le \alpha} Z_\beta \otimes Z_{\alpha - \beta}$$

Then $V_p(Z_\alpha) = Z_{\alpha/p}$, where $Z_{\alpha/p}$ means 0 if p does not divided α.

Proof. If the Z_α's are linearly independent and span H, or in other words if they form a structural basis, then the result follows from [A, theorem 2.5.9]. In general we can find a coalgebra with a structural basis together with a homomorphism mapping the elements of the structural basis to the Z_α's. The result then follows because the shifts V_p commute with homomorphisms. This proves lemma 2.11.

Theorem 2.12. *Suppose that H is a $\mathbf{Z}_{\geq 0}$-graded cocommutative bialgebra over $\mathbf{Z}$ whose homogeneous pieces are finitely generated free abelian groups and such that the degree 0 piece is spanned by 1. Suppose that the shifts $V_p : H/pH \to H/pH$ are onto for all primes p. Then every primitive element of H is liftable.*

Remark. In fact theorem 2.15 (whose proof uses theorem 2.12) implies the stronger result that H has a structural basis.

Proof. We first show that if K is any field then the K bialgebra $C = H \otimes K$ is irreducible, in other words it has only one simple subcoalgebra. By theorem 2.3.4 of [A], the simple subcoalgebras of a K-coalgebra C correspond to the maximal ideals of the dual algebra C^* which are not dense, so it is sufficient to show that C^* is a local ring. The elements of C^* are infinite sums of the form $c = c_0 + c_1 + \cdots$ with $c_i \in C_i^*$ (where C_i is the degree i subspace of C). We show that the ideal M of all elements with $c_0 = 0$ is the unique maximal ideal of C^*, in other words that any element c with $c_0 \neq 0$ invertible. The space C_0^* is isomorphic to the field K, so if $c_0 \neq 0$ then it is invertible, so we can recursively define an inverse $d = d_0 + d_1 + \cdots$ to c by putting $d_0 = c_0^{-1}$, $d_n = -c_0^{-1}(d_{n-1}c_1 + \cdots + d_0 c_n)$ for $n > 0$. This shows that C^* is a local ring, so C is irreducible.

Next we check that $H \otimes \mathbf{F}_p$ and $H \otimes \mathbf{Q}$ have structural bases. By [A, theorem 2.4.24] any irreducible bialgebra over a field, in particular $H \otimes \mathbf{F}_p$, is automatically a Hopf algebra. By [A,corollary 2.5.15] any cocommutative irreducible Hopf algebra over $\mathbf{F}_p$ such that V_p is onto, in particular $H \otimes \mathbf{F}_p$, has a structural basis. Similarly theorem 2.5.3 of [A] states that any irreducible cocommutative Hopf algebra over a field of characteristic 0 is a universal enveloping algebra of its primitive elements. So $H \otimes \mathbf{Q}$ is the universal enveloping algebra of its primitive elements and therefore also has a structural basis.

The next step is to show that any primitive element of $H \otimes \mathbf{F}_p$ can be lifted to a primitive element of H. If m_α is the dimension of the space of primitive elements of degree α of a $\mathbf{Z}_{\geq 0}^n$-graded Hopf algebra with a structural basis and n_α is the dimension of the space of all elements of degree α, then

$$\prod_{\alpha > 0} (1 - x^\alpha)^{-m_\alpha} = \sum_{\alpha \geq 0} n_\alpha x^\alpha$$

by lemma 2.4. In particular the spaces of primitive elements of degrees α in $H \otimes \mathbf{F}_p$ and $H \otimes \mathbf{Q}$ have the same dimension, because it is obvious that the spaces of all elements in $H \otimes \mathbf{F}_p$ and $H \otimes \mathbf{Q}$ of degree α have the same dimension, and $H \otimes \mathbf{F}_p$ and $H \otimes \mathbf{Q}$ both have structural bases by the paragraph above. The space P_α of primitive elements of H of degree α also has the same rank as the space of primitive elements of $H \otimes \mathbf{Q}$, and therefore the same rank as the space $P_{\alpha,p}$ of primitive elements of $H \otimes \mathbf{F}_p$ of degree α. The map from P_α/pP_α to $P_{\alpha,p}$ is obviously injective and both spaces are finite dimensional vector spaces over $\mathbf{F}_p$ of the same dimension, so the map between them is also surjective, or in

other words every primitive element of $H \otimes \mathbf{F}_p$ (of any degree α) can be lifted to a primitive element of H. By lemma 2.10 this implies every primitive element of H is liftable. This proves theorem 2.12.

Corollary 2.13. *Let k and n be positive integers. If*

$$a(x) = \sum_{i_1,\dots,i_k \geq 0} a_{i_1,\dots,i_k} x_1^{i_1} \cdots x_k^{i_k}$$

is group-like, for $a_{i_1,\dots,i_k}$ in some bialgebra over a commutative ring R, and $a_{0,\dots,0} = 1$, $a_{i_1,\dots,i_k} = 0$ for $0 < i_1 + \cdots + i_k < n$, then $a_{i_1,\dots,i_k}$ is liftable for $i_1 + \cdots + i_k = n$.

Proof. It is sufficient to prove this for the universal example, where the bialgebra H is the free associative algebra over $\mathbf{Z}$ generated by $a_{i_1,\dots,i_k}$ for $i_1 + \cdots, i_k \geq n$ and the coalgebra structure is defined so that $a(x)$ is group-like. Put $a_0 = 1$, $a_{i_1,\dots,i_k} = 0$ for $0 < i_1 + \cdots i_k < n$, and grade H by letting $a_{i_1,\dots,i_k}$ have degree $i_1 + \cdots + i_k$. Then $V_p(a_{i_1\dots,i_k}) = a_{i_1/p,\dots,i_k/p}$ by lemma 2.11, so $V_p : H/pH \mapsto H/pH$ is onto. By theorem 2.12 this implies that any primitive element of H, in particular $a_{i_1,\dots,i_k}$ for $i_1 + \cdots + i_k < 2n$, is liftable. This proves corollary 2.13.

Theorem 2.14. *The set of liftable elements of a cocommutative bialgebra H over a commutative ring R is a Lie algebra over R.*

Proof. We already know that the liftable primitive elements form a module over R by lemma 2.5. We have to show that the set of liftable elements is closed under the Lie bracket. Suppose a_1 and b_1 are two primitive elements with group-like lifts $a(x) = \sum a_i x^i$ and $b(y) = \sum_i b_i y^i$. It is sufficient to prove theorem 2.14 in the case of the universal example, when H is the free associative algebra over $\mathbf{Z}$ generated by the elements a_i and b_i for $i > 0$, with the coalgebra structure defined by the fact that $a(x)$ and $b(y)$ are group-like. We $\mathbf{Z}$-grade H by giving a_i and b_i degree $i \in \mathbf{Z}$. By lemma 2.11 the shifts V_p contain all the elements a_i and b_i in their images, so the shifts are onto as they are homomorphisms of $\mathbf{F}_p$-algebras. Then by theorem 2.12 we see that all primitive elements of H, in particular $[a_1, b_1]$, are liftable. This proves theorem 2.14.

Open problem: give a direct proof of theorem 2.14 by writing down an explicit lifting of $[a_1, b_1]$ as a non-commutative polynomial in the a_i's and the b_i's. This seems surprisingly difficult, possibly because such a lifting is far from unique so it is hard to think of a canonical way to define it.

Theorem 2.15. *Suppose that U is a torsion free bialgebra over a principal ideal domain R, and suppose that the Lie algebra of primitive elements of U is a free R-module. If U is generated as an algebra by all coefficients a_i of some set of group-like elements $1 + a_1 x + a_2 x^2 + \cdots \in U[[x]]$ then U is a Hopf algebra with a structural basis.*

Proof. As R is a principal ideal domain, any submodule of a free module is free, and in particular the Lie algebra L of liftable primitive elements of U is a free R module. Suppose that I is a totally ordered set indexing a basis $a_1^i, i \in I$ of L. (The superscript on a_1^i is just used as an index and has nothing to do with powers of a_1.) For each a_1^i chose

a lifting $a^i(x) = \sum_n a^i_n(x)$. Define H to be the R-module spanned by the elements of the form $a^{i_1}_{n_1} \cdots a^{i_k}_{n_k}$ with $i_1 < i_2 \cdots < i_k$.

We need to check that the natural map from $H \otimes_R H$ to $U \otimes_R U$ is injective, so that $H \otimes_R H$ can be regarded as a submodule of $U \otimes_R U$. As U is torsion-free, its submodule H is also torsion free and therefore flat (as R is a principal ideal domain). Hence $H \otimes_R H \subseteq U \otimes_R U$ (as $H \otimes_R H \subseteq U \otimes_R H \subseteq U \otimes_R U$).

The spanning set of H of elements of the form $a^{i_1}_{n_1} \cdots a^{i_k}_{n_k}$ satisfies the relations of a structural basis, so in particular H is a sub coalgebra of U. This set of elements is linearly independent by lemma 2.7 because the primitive elements a^i_1 in it are linearly independent, so it forms a structural basis for the coalgebra H. We will show that U has a structural basis by showing that it is equal to H. To do this we will first show that H is closed under multiplication, and then show that H contains a set of generators of the algebra U.

We define a filtration of H by letting H_n be the subspace of H spanned by all the monomials $a^{i_1}_{n_1} \cdots a^{i_k}_{n_k}$ with $n_1 + \cdots n_k \leq n$. We will show by induction on N that $H_m H_n \subseteq H_{m+n}$ whenever $m + n \leq N$; this is trivial for $N = 1$. To do this it is sufficient to show that for any r, s with $r + s \leq N$ we have $a^i_r a^j_s \equiv a^j_s a^i_r \bmod H_{r+s-1}$ whenever $i > j$, and $a^i_r a^i_s \equiv \binom{r+s}{r} a^i_{r+s} \bmod H_{r+s-1}$.

We will show by induction on k that if $1 \leq k \leq N$ then there is a group-like element $c(x, y) = \sum c_{ij} x^i y^j$ such that $c_{ij} \in H_{i+j}$ if $i + j \leq N$ and $c_{ij} = b_j a_i$ for $i + j \leq k$. This is obvious for $k = 1$, as we can take $c_{ij} = a_i b_j$. Suppose we have proved this for some value of $k < N$, so we wish to prove it for $k+1$. Consider the group-like element $b(y)a(x)c(x, y)^{-1}$. The coefficients of $x^i y^j$ of this are 0 for $0 < i + j \leq k$, so by corollary 2.13 the coefficients for $i + j = k + 1$ are all liftable. For each of them choose a lifting $d_i(z)$, such that the coefficient of x^j of any of these liftings is in H_j. We can construct such liftings as products $a^{i_1}(r_1 x)a^{i_2}(r_2 x) \cdots$ if $d_i = r_1 a^{i_1} + r_2 a^{i_2} + \cdots$ with $i_1 < i_2 < \cdots$. Then the product

$$e(x, y) = \Big(\prod_{1 \leq i \leq k} d_i(x^i y^{k+1-i}) \Big) c(x, y)$$

satisfies the conditions above for $k + 1$. In fact by construction we see that $e_{ij} = b_j a_i$ for $i + j \leq k + 1$. Also the coefficient of $x^i y^j$ in $e(x, y)$ is either already an element of H_{i+j} or it is a sum of products of elements in $H_{i_1}, H_{i_2}, \ldots$ with $i_1 + i_2 + \cdots < i + j$. If $i + j \leq N$ then by the inductive assumption on N such product is already in $H_{i+j-1} \subseteq H_{i+j}$, so the coefficient of $x^i y^j$ is in H_{i+j} provided $i + j \leq N$. This proves the inductive hypothesis for all $k \leq N$.

If we take $k = N$ in the inductive hypothesis above, we now see that $b_j a_i \in H_{i+j}$ for $i + j = N$. By induction on N we see that $b_j a_i \in H_{i+j}$ for all i, j.

The proof that $a_i a_j = \binom{i+j}{j} a_{i+j} \bmod H_{i+j-1}$ is similar, except that we use the group-like element $a(x + y) = \sum_{i,j} \binom{i+j}{i} a_{i+j} x^i y^j$ instead of $b(x)a(y)$.

This shows that H is a subalgebra of U (and hence a Hopf algebra with a structural basis).

Finally we check that H contains the coefficients of any lifting $a(x)$ of any primitive element a_1 of U. We will show by induction on n that if $n \geq 1$ then there is a lifting $b(x)$ of a_1 such that all coefficients of $b(x)$ are in H and $b_i = a_i$ for $i \leq n$. This is clear for $n = 1$ as we just take any lifting of a_1 with coefficients in H. Suppose that $b_i = a_i$ for

$i \le n$. Then $b(x)^{-1}a(x)$ has all its coefficients of x^i vanishing for $0 < i \le n$, so by corollary 2.13 (with $k = 1$) we see that the coefficient of x^{n+1} is liftable. Choose a lifting $c(x)$ with coefficients in H. Then $b(x)c(x^{n+1})$ has all coefficients in H and the coefficients of x^i for $i \le n+1$ are equal to a_i. This proves the inductive hypothesis for $n+1$, and hence for all n. Therefore all coefficients of $a(x)$ are in H.

We have shown that H is a bialgebra with a structural basis and that H contains a set of generators of the algebra U. Therefore H is equal to U, and U has a structural basis. This proves theorem 2.15.

Example 2.16. Suppose n is a nonnegative integer. The $\mathbf{Z}$-Hopf algebra F_n is defined to the free associative algebra over $\mathbf{Z}$ on a set of elements Z_α, $\alpha \in \mathbf{Z}^n_{\ge 0}$, $\alpha \ne 0$, with the comultiplication defined by

$$\Delta(Z_\alpha) = \sum_{0 \le \beta \le \alpha} Z_\beta \otimes Z_{\alpha-\beta}$$

on the generators (where $Z_0 = 1$). Then it follows immediately from theorem 2.15 that F_n has a structural basis. The fact that $F_n \otimes \mathbf{F}_p$ has a structural basis was first proved by Dieudonné; more precisely it follows from the comments at the end of section 15 of [D], which show that $F_n(\mathbf{F}_p)$ has a structural basis, together with theorem 3 of [D], which shows that Dieudonné's definition of $F_n(\mathbf{F}_p)$ is equivalent to $F_n \otimes \mathbf{F}_p$. There are proofs that F_1 has a structural basis in [D72, proposition 2.2] and [Sh], but unfortunately the paper [Sh] seems to be unpublished, and the proof in [D72] has a gap (see [H p. 516]): in [D72, page 5, line 6], it is implicitly assumed that I_k/I_k^2 is torsion free. (It follows from example 2.16 that this is indeed true, but it seems rather hard to prove.)

Example 2.17. Suppose that U is a torsion free bialgebra over a principal ideal domain R, and suppose that the Lie algebra of primitive elements of U is a free R-module. Then U has a subalgebra H with a structural basis containing all subalgebras with a structural base. The primitive elements of H are exactly the liftable primitive elements of U, and H is generated as an algebra by all coefficients of all liftings of primitive elements of U. This follows easily from theorem 2.15.

3. Liftings of Lie algebra elements.

In this section we construct liftings of certain elements of the Lie algebra of the vertex algebra of a double cover of an even lattice ([B86]). More precisely we construct liftings of elements in the root spaces of vectors of norms 2 or 0. This will be enough for applications to the fake monster Lie algebra, because its rational form is generated by the root spaces of roots of norms 2 or 0. For vectors in norm 2 root spaces we construct a formal one parameter group lifting any element in the root space. For norm 0 roots we cannot do this, but have to make do with lifting vectors to formal curves.

Lemma 3.1. *Suppose a is an element of a vertex algebra, and define a^k for $k \ge 0$ by $a^0 = 1$, $a^{k+1} = a_{-1}(a^k)$. Then*

$$a_n^k(b) = \sum_{0 \le j \le k} \binom{k}{j} \sum_{\substack{i_1,\dots,i_j < 0 \\ i_{j+1},\dots,i_k \ge 0 \\ i_1+\cdots+i_k = n-k}} a_{i_1} \cdots a_{i_k} b$$

Proof. This follows by induction on n using the fact that

$$(a_{-1}a^k)_n b = \sum_{i\geq 0} a_{-1-i}(a^k_{n+i}b) + \sum_{i\geq 0} a^k_{n-1-i}(a_i b)$$

which is a special case of the vertex algebra identity. This proves lemma 3.1.

Corollary 3.2. *Suppose that a is an element of a vertex algebra V such that $a_i a = 0$ for $i \geq -1$. Then for each $k \geq 0$ the operator a_0^k is divisible by $k!$.*

Proof. We show that a_0^k is divisible by $k!$ by induction on k, which is obvious for $k = 0, 1$. As $a_{-1}a = 0$ the left hand side of lemma 3.1 vanishes for $k \geq 2$. As $a_i a = 0$ for $i \geq 0$ all the operators a_i commute with each other because of the formula

$$[a_i, b_j] = \sum_{k\geq 0} \binom{i}{k} (a_k b)_{i+j-k},$$

which together with the induction hypothesis implies that the sum of all terms on the right hand side of lemma 3.1 other than a_0^k is divisible by $k!$. Hence a_0^k is also divisible by $k!$. This proves lemma 3.2.

Corollary 3.3. *Suppose V is the vertex algebra (over $\mathbf{Z}$) of a double cover $\hat{L}$ of an even integral lattice L. If $\alpha \in \hat{L}$ has positive norm and a is the element e^α of V, then all coefficients of*

$$\exp(xa_0) = \sum_k x^k a_0^k/k!$$

map V into V.

Proof. For any $\alpha, \beta \in L$ we have $e_i^\alpha e^\beta = 0$ for $i + (\alpha, \beta) \geq 0$. In particular $e_i^\alpha e^\alpha = 0$ for $i \geq -1$ if $(\alpha, \alpha) > 0$. Corollary 3.3 now follows from corollary 3.2.

Corollary 3.3 will allow us to lift elements in the root spaces of norm 2 vectors. In the rest of this section we show how to lift elements in the root spaces of norm 0 vectors.

Lemma 3.4. *Suppose that the the elements Γ_i for $i \in \mathbf{Z}$ are independent formal variables. If $I = (i_1, \ldots, i_m)$ is a finite sequence of integers then define Γ_I to be $\Gamma_{i_1} \cdots \Gamma_{i_m}$ and define $\Sigma(I)$ to be $i_1 + \cdots + i_m$ and define $l(I)$ to be m. Then all coefficient of the power series*

$$E = \exp\Big(\sum_{m,n>0} \sum_{I\in\mathbf{Z}^m} \sum_{\substack{J\in\mathbf{Z}^n \\ \Sigma(J)=-\Sigma(I)>0}} \frac{\Sigma(I)}{mn} \Gamma_I \Gamma_J \Big)$$

are integers.

Proof. If I is any finite sequence of integers then I^k means the obvious concatenation of k copies of the sequence I. We will call a pair of finite sequences primitive if it is not of the form (I^k, J^k) for some $k \geq 2$. Any pair of finite sequences can be written uniquely as (I^k, J^k) for some primitive pair (I, J), which we call the primitive core of the pair (I^k, J^k). Consider the group $\mathbf{Z} \times \mathbf{Z}$ acting on pairs (I, J) by the first $\mathbf{Z}$ acting as

cyclic permutations of the elements of I and the second $\mathbf{Z}$ acting as cyclic permutations of the elements of J. We group the pairs (I,J) indexing the terms of the sum in the exponent of E into equivalence classes, where we say two pairs of sequences are equivalent if their primitive cores are conjugate under $\mathbf{Z}\times\mathbf{Z}$. For any primitive element (I,J) with $\Sigma(I)+\Sigma(J)=0$ we will show that all coefficients of the exponential of the sum of all terms in the equivalence class of (I,J) are integral. This will show that the coefficients of E are integral because E is an infinite product over the set of orbits of primitive elements of expressions like this. Let m' and n' be the number of orbits of I and J under the cyclic action of $\mathbf{Z}$. Then the exponential of the sum of the terms equivalent to (I,J) is

$$\begin{aligned}
&\exp\Big(\sum_{k>0}\frac{\Sigma(I^k)}{l(I^k)l(J^k)}(\text{number of orbits of } (I,J) \text{ under } \mathbf{Z}\times\mathbf{Z})\Gamma_{I^k}\Gamma_{J^k}\Big)\\
=&\exp\Big(\sum_{k>0}\frac{k\Sigma(I)m'n'}{kmkn}\Gamma_{I^k}\Gamma_{J^k}\Big)\\
=&\exp\Big(\frac{\Sigma(I)m'n'}{mn}\sum_{k>0}\Gamma_I^k\Gamma_J^k/k\Big)\\
=&\exp\Big(-\frac{\Sigma(I)m'n'}{mn}\log(1-\Gamma_I\Gamma_J)\Big)\\
=&(1-\Gamma_I\Gamma_J)^{-\Sigma(I)m'n'/mn}
\end{aligned}$$

The number $\Sigma(I)$ is divisible by m/m' as I is the concatenation of m/m' identical sequences and $\Sigma(I)$ is the sum of the elements of I, and similarly $\Sigma(I)=-\Sigma(J)$ is divisible by n/n'. Moreover m/m' and n/n' are coprime as (I,J) is primitive. Hence $\Sigma(I)m'n'/mn$ is an integer because it is equal to $\Sigma(I)$ divided by two coprime factors m/m' and n/n' of $\Sigma(I)$. This implies that $(1-\Gamma_I\Gamma_J)^{-\Sigma(I)m'n'/mn}$ has integral coefficients, and so E does as well as it is an infinite product of expressions like this. This proves lemma 3.4.

Theorem 3.5. *Let V be the vertex algebra (over the integers) of the double cover of an even lattice L. Define D^*V to be the sum of the spaces $D^{(i)}V$ for $i\geq 1$. We recall from [B86] that V/D^*V has a natural Lie algebra structure, with the bracket defined by $[u,v]=u_0v$. Moreover this Lie algebra acts on V preserving the vertex algebra structure of V. Suppose that $\alpha,\gamma\in L$ with α orthogonal to the norm 0 vector γ. Then*

$$\exp\Big(\sum_{i>0}\frac{x^i}{i}(\alpha(1)e^{i\gamma})_0\Big)$$

is a lifting of $(\alpha(1)e^{\gamma})_0$ in the universal enveloping algebra of $(V/D^*V)\otimes\mathbf{Q}$ all of whose coefficients map V to V.

Proof. It is obvious that the element is group-like as it is the exponential of a primitive element, so the only problem is to show that it preserves the integral form V.

This element is also an automorphism of the vertex algebra $(V\otimes\mathbf{Q})[[x]]$ as it is the exponential of a derivation of this vertex algebra. Hence to show its coefficients preserve

the integral form of V it is sufficient to show that it maps each of the generators e^β, $\beta \in L$, of the vertex algebra V into $V[[x]]$.

Define the operators Γ_i by

$$e^\gamma(z) = \sum_i \Gamma_i z^i.$$

Recall the following formulas from [B86].

$$\alpha(1)(z) = \sum_j \alpha(1)_{-j} z^{j-1} = \sum_j \alpha(j) z^{j-1}$$

All coefficients of e^γ commute with everything in sight as γ is orthogonal to γ and α. So $(\alpha(1)e^{i\gamma})_0$ is equal to the coefficient of z^{-1} in

$$\alpha(1)(z)(e^\gamma(z))^i = \sum_j \alpha(j) z^{j-1} \sum_{I \in \mathbf{Z}^i} \Gamma_I z^{\Sigma(I)}$$

Therefore $\sum_{i>0} \frac{x^i}{i}(\alpha(1)e^{i\gamma})_0$ is equal to $A^+ + A^0 + A^-$ where

$$A^+ = \sum_{m>0} \sum_{j>0} \alpha(j) \sum_{\substack{I \in \mathbf{Z}^m \\ \Sigma(I) = -j}} \frac{x^m}{m} \Gamma_I$$

$$A^0 = \sum_{m>0} \sum_{j=0} \alpha(j) \sum_{\substack{I \in \mathbf{Z}^m \\ \Sigma(I) = -j}} \frac{x^m}{m} \Gamma_I$$

$$A^- = \sum_{m>0} \sum_{j<0} \alpha(j) \sum_{\substack{I \in \mathbf{Z}^m \\ \Sigma(I) = -j}} \frac{x^m}{m} \Gamma_I$$

We would like to pull out a factor of $\exp(A^-)$ from $\exp(A^+ + A^0 + A^-)$ because $\exp(A^-)(e^\beta) = e^\beta$, but we have to be careful when doing this because $\alpha(i)$ does not commute with $\alpha(-i)$ if $i \neq 0$.

We can easily evaluate $[A^+, A^-]/2$ using the fact that $[\alpha(i), \alpha(j)] = j(\alpha, \alpha)$ if $i+j = 0$ and 0 otherwise, and we find that

$$[A^+, A^-]/2 = \frac{1}{2}(\alpha, \alpha) \sum_{m,n>0} \sum_{k<0} \sum_{\substack{I \in \mathbf{Z}^m \\ \Sigma(I) = k}} \sum_{\substack{J \in \mathbf{Z}^n \\ \Sigma(J) = -k}} \frac{k}{mn} \Gamma_I \Gamma_J.$$

In particular this verifies the fact used below that $[A^+, A^-]$ commutes with A^+ and A^-. The exponential of this is the expression in lemma 3.4 raised to the power of $(\alpha, \alpha)/2$, and therefore has integral coefficients by lemma 3.4 and because (α, α) has even norm. So $\exp([A^+, A^-]/2)$ maps V into $V[[x]]$.

Recall the formula

$$\exp(A^+ + A^-) = \exp([A^-, A^+]/2) \exp(A^+) \exp(A^-)$$

which is valid because $[A^+, A^-]$ commutes with A^+ and A^-. (This is essentially just the first few terms of the Baker-Campbell-Hausdorff formula.) Now we look at

$$\exp(A^+ + A^0 + A^-)e^\beta = \exp([A^-, A^+]/2)\exp(A^+)\exp(A^0)\exp(A^-)e^\beta$$

It is obvious that $\exp(A^-)(e^\beta) = e^\beta$, and we have checked above that all coefficients of $\exp([A^-, A^+]/2)$ map V to V. Hence to complete the proof of theorem 3.5 it is sufficient to prove that all coefficients of $\exp(A^0)$ and $\exp(A^+)$ map V into V.

We check that $\exp(A^0)$ maps V to $V[[x]]$. The follows because A^0 is an infinite sum of expressions like

$$\sum_{n>0} \alpha(0)\Gamma_{J^n}x^n/n = -\alpha(0)\log(1 - x\Gamma_J)$$

where J is a primitive sequence (in other words a sequence that cannot be written in the form I^m for some $m > 1$). Hence $\exp(A^0)$ is an infinite product of terms of the form $(1 - x\Gamma_J)^{-\alpha(0)}$, which map $V[[x]]$ to itself because $\alpha(0)$ has integral eigenvalue (α, β) on the subspace of V of degree $\beta \in L$. This shows that all coefficients of $\exp(A^0)$ map V into V.

Finally we have to show that all coefficients of $\exp(A^+)$ map V into V. As usual we divide the sum over elements I in A^+ into classes consisting of powers of conjugates of primitive elements I. We see that A^+ is a sum over all orbits of primitive elements I with $\Sigma(I) > 0$ of expressions like

$$l(I)\sum_{k>0} \alpha(\Sigma(I^k))\frac{x^{l(I^k)}}{l(I^k)}\Gamma_{I^k} = \sum_{k>0} \alpha(k\Sigma(I))\frac{x^{kl(I)}}{k}\Gamma_I^k.$$

(The factor at the front is the number of conjugates of I under the cyclic action of $\mathbf{Z}$, which is equal to $l(I)$ because I is primitive.) So it is sufficient to show that the exponential of this expression has integral coefficients. Let $y_1, y_2, \ldots,$ be a countable number of independent variables, and identify $\alpha(k)$ with the symmetric function $\sum y_i^k$ of the y's for $k > 0$. (See [M chapter 1].) Then

$$\begin{aligned}
&\exp\Big(\sum_{k>0} \alpha(k\Sigma(I))\frac{x^{kl(I)}}{k}\Gamma_I^k\Big)\\
=&\exp\Big(\sum_i \sum_{k>0} y_i^{k\Sigma(I)}\frac{x^{kl(I)}}{k}\Gamma_I^k\Big)\\
=&\exp\Big(\sum_i -\log(1 - y_i^{\Sigma(I)}x^{l(I)}\Gamma_I)\Big)\\
=&\prod_i \frac{1}{1 - y_i^{\Sigma(I)}x^{l(I)}\Gamma_I}.
\end{aligned}$$

So we see that

$$\exp(A^+) = \prod_I \prod_i \frac{1}{1 - y_i^{\Sigma(I)}x^{l(I)}\Gamma_I}$$

where the product over I is a product over all orbits of primitive sequences I with $\Sigma(I) > 0$ under the cyclic action of $\mathbf{Z}$. The last line is a power series in the elements Γ_I and x whose coefficients are symmetric functions in the y's, and hence are polynomials with integral coefficients in the complete symmetric functions of the y's. So we have to show that each complete symmetric function, considered as a polynomial in the α's with rational (not necessarily integral!) coefficients maps V to V. The complete symmetric functions are the polynomials $e^{-\alpha}D^{(n)}(e^{\alpha})$ considered as elements of the ring V underlying the vertex algebra V (as follows from [M, Chapter 1, 2.10]). By definition of the integral form V these polynomials map V to itself.

This proves theorem 3.5.

4. The fake monster smooth Hopf algebra.

We recall the construction of the fake monster Lie algebra [B90]. It is the Lie algebra of physical states of the vertex algebra of the double cover of the lattice $II_{25,1}$. This Lie algebra has an integral form $\mathbf{m}$ consisting of the elements represented by elements of the integral form of the vertex algebra. We recall some properties of $\mathbf{m}$:

1 $\mathbf{m}$ is graded by the lattice $II_{25,1}$, and the piece $\mathbf{m}_\alpha$ of degree $\alpha \in II_{25,1}$ has dimension $p_{24}(1-\alpha^2/2)$ if $\alpha \neq 0$ and 26 if $\alpha = 0$, where $p_{24}(n)$ is the number of partitions of n into parts of 24 colors.

2 $\mathbf{m}$ has an involution ω lifting the involution -1 of $II_{25,1}$.

3 $\mathbf{m}$ has a symmetric invariant integer valued bilinear form $(,)$, and the pairing between $\mathbf{m}_\alpha \otimes \mathbf{Q}$ and $\mathbf{m}_{-\alpha} \otimes \mathbf{Q}$ is nonsingular.

4 $\mathbf{m} \otimes \mathbf{Q}$ is a generalized Kac-Moody algebra. The simple roots are given by the norm 2 vectors r with $(r,\rho) = -1$, together with all positive multiples of ρ with multiplicity 24, where ρ is a primitive norm 0 vector of $II_{25,1}$ such that $\rho^\perp/\rho$ is isomorphic to the Leech lattice. (This follows from [B90. theorem 1].)

We define $U^+(\mathbf{m})$ to be the $\mathbf{Z}$-subalgebra of the universal enveloping algebra $U(\mathbf{m}\otimes\mathbf{Q})$ generated by the coefficients of the liftings of elements in root spaces of the simple roots and their negatives constructed in corollary 3.3 and theorem 3.5. (Note that all simple roots of $\mathbf{m}$ have norms 2 or 0 so we can always apply one of these two types of liftings.)

Theorem 4.1. *There is a $II_{25,1}$-graded Hopf algebra $U^+(\mathbf{m})$ over $\mathbf{Z}$ with the following properties.*

1 $U^+(\mathbf{m})$ has a structural basis over $\mathbf{Z}$.

2 The primitive elements of $U^+(\mathbf{m})$ are an integral form of the fake monster Lie algebra $\mathbf{m}$.

3 For every norm 2 vector of $II_{25,1}$, $U^+(\mathbf{m})$ contains the usual (Kostant) integral form of the universal enveloping algebra of the corresponding $sl_2(\mathbf{Z})$.

Proof. The algebra $U^+(\mathbf{m})$ is a $\mathbf{Z}$-Hopf subalgebra of $U(\mathbf{m} \otimes \mathbf{Q})$ as it is generated by coalgebras. Also $U^+(\mathbf{m})$ is obviously torsion free as it is contained in a rational vector space. It is easy to check directly that the degree zero primitive elements of $U^+(\mathbf{m})$ are just the degree 0 elements of $\mathbf{m}$ and therefore form a free $\mathbf{Z}$ module (of rank 26). If α is any nonzero element of $II_{25,1}$ having nonzero inner product with some element $\beta \in II_{25,1}$, then $(\alpha,\beta)u \in \mathbf{m}_\alpha$ for any primitive element $u \in U^+(\mathbf{m})$ of degree α, because $U^+(\mathbf{m})$ maps $\mathbf{m}$, and hence β, to $\mathbf{m}$ by corollary 3.3 and theorem 3.5. This shows that all root

spaces of primitive elements of $U^+(\mathbf{m})$ are free $\mathbf{Z}$-modules. We can now apply theorem 2.15 to see that $U^+(\mathbf{m})$ is a Hopf algebra with a structural basis. The primitive elements of $U^+(\mathbf{m})$ form an integral form of the fake monster Lie algebra, because the fake monster Lie algebra over the rationals is generated by the root spaces of simple roots and their negatives, and all simple roots have norms 2 or 0. This proves theorem 4.1.

5. A smooth Hopf algebra for the Virasoro algebra.

In this section we show that there is a Hopf algebra over $\mathbf{Z}$ with a structural basis whose primitive elements form the natural integral form of the Virasoro algebra (theorem 5.7). In other words, there is a formal group law over $\mathbf{Z}$ corresponding to the Virasoro algebra. Moreover this Hopf algebra acts on the integral form of the vertex algebra of any even self dual lattice.

Let R be a commutative ring. We write $Hom(R,R)$ for the ring of homomorphisms of the abelian group R to itself, and $Der(R)$ for the Lie algebra of derivations of the ring R, and $U(Der(R))$ for the universal enveloping algebra of $Der(R)$ over $\mathbf{Z}$. Consider the group of all element $a = \sum_i a_i \epsilon^i \in Hom(R,R)[[\epsilon]]$ with $a_0 = 1$ that induce automorphisms of the $\mathbf{Z}[[\epsilon]]$ algebra $R[[\epsilon]]$. We can think of the elements of this group informally as "infinitesimal curves in the group of automorphisms of $Spec(R)$". We will call a derivation of R liftable if it is of the form a_1 for some a as above.

Lemma 5.1. *Let R be any commutative algebra with no $\mathbf{Z}$-torsion. Then any $a \in Hom(R,R)[[\epsilon]]$ with $a(0) = 1$ that is an automorphism of $R[[\epsilon]]$ is the image of a unique group-like element G_a of $U(Der(R) \otimes \mathbf{Q})[[\epsilon]]$ with $G_a(0) = 1$ under the natural map from $U(Der(R) \otimes \mathbf{Q})$ to $Hom(R,R) \otimes \mathbf{Q}$.*

Proof. We note that $\log(a)$ is a well defined element of $(Hom(R,R) \otimes \mathbf{Q})[[\epsilon]]$ as $a = 1 + O(\epsilon) \in Hom(R,R)[[\epsilon]]$. As a is a ring homomorphism it follows that that $\log(a)$ is a derivation of $(R \otimes \mathbf{Q})[[\epsilon]]$ and is therefore an element of $(Der(R) \otimes \mathbf{Q})[[\epsilon]]$. Now we consider $\log(a)$ to be an element of the universal enveloping algebra $U(Der(R) \otimes \mathbf{Q})[[\epsilon]]$ and we define $G_a \in U(Der(R) \otimes \mathbf{Q})[[\epsilon]]$ by

$$G_a = \exp(\log(a)),$$

where the log is computed in $(Hom(R,R) \otimes \mathbf{Q})[[\epsilon]]$ and the exponential is computed in $U(Der(R) \otimes \mathbf{Q})[[\epsilon]]$. It is obvious that the action of G_a on $R[[\epsilon]]$ is the same as that of a. Also G_a is group-like because it is the exponential of a primitive element. It is easy to check that G_a is the unique group-like lifting of a with $G_a(0) = 1$, because the log of a group-like element must be primitive and must therefore be the same as $\log(a)$. This proves lemma 5.1.

Corollary 5.2. *Suppose that R is a commutative ring with no $\mathbf{Z}$ torsion such that $Der(R)$ is a free $\mathbf{Z}$-module. Define $U^+(Der(R))$ to be the subalgebra of $U(Der(R \otimes \mathbf{Q}))$ generated by all the coefficients of all group-like elements of $U(Der(R) \otimes \mathbf{Q})[[\epsilon]]$ that have constant coefficient 1 and map $R[[\epsilon]]$ to $R[[\epsilon]]$. Then $U^+(Der(R))$ is a Hopf algebra over $\mathbf{Z}$ with a structural basis, and its primitive elements are the liftable primitive elements of $Der(R)$.*

Proof. Applying theorem 2.15 shows that $U^+(Der(R))$ is a Hopf algebra over $\mathbf{Z}$ with a structural basis. By lemma 5.1 the space of primitive elements of $U^+(Der(R))$ is the same as the space of liftable primitive elements of $Der(R)$. This proves corollary 5.2.

Example 5.3 Suppose we take R to be the algebra $\mathbf{Z}[x][x^{-1}]$ of Laurent polynomials. Then $Der(R) = Witt$ is the Witt algebra over $\mathbf{Z}$, which is spanned by the elements $L_m = -x^{m+1}\frac{d}{dx}$ for $m \in \mathbf{Z}$. All elements L_m are liftable; for example, we can use the automorphism of $R[[\epsilon]]$ taking x to $x - \epsilon x^{m+1}$ to show that L_m is liftable. Therefore the Hopf algebra $U^+(Witt)$ is a Hopf algebra over $\mathbf{Z}$ with a structural basis, whose primitive elements are exactly the elements of the Witt algebra.

Lemma 5.4. *Let N be an integer. Let R_N be the representation of $Witt$ with a basis of elements e_n, $n \in \mathbf{Z}$, with the action given by $L_m(e_n) = (Nm+n)e_{m+n}$. Then the action of $Witt$ on R_N can be extended to an action of $U^+(Witt)$ on R_N.*

Proof. For $N = -1$ the module R_N is the module of first order differential operators on $R = \mathbf{Z}[x][x^{-1}]$, so the automorphisms G_a extend to $R_N[[\epsilon]]$. For other negative values of N the R-module R_N is a tensor product of $-N$ copies of the module R_{-1}, and R_N for N positive is the dual of R_N for N negative. Therefore $U^+(Witt)$ extends to these modules as well. This proves lemma 5.4.

We define $Witt_{\geq n}$ for $n = -1, 1$ to be the subalgebra of $Witt$ spanned by L_i for $i \geq n$. Let V be the vertex algebra of some even lattice. It contains elements e^α for $\alpha \in L$, so there are operators e^α_i on V for $i \in \mathbf{Z}$. The algebra $Witt_{\geq -1}$ also acts naturally on V.

Lemma 5.5. *Put $N = \alpha^2/2 - 1$ and $e_j = e^\alpha_{\alpha^2/2-1-j}$ for $j \in \mathbf{Z}$ and let R_N be the space with the elements e_j as a basis. Define an action of the algebra $U^+(Witt_{\geq -1})$ on $R_N \otimes U^+(Witt_{\geq -1})$ using the action on R_N as in lemma 5.4 and the action on $U^+(Witt_{\geq -1})$ by left multiplication and the coalgebra structure of $U^+(Witt_{\geq -1})$. If $u \in U^+(Witt_{\geq -1})$ then $ue^\alpha_m = \sum_i e^\alpha_i u_i$ as operators on V, where $\sum_i e^\alpha_i \otimes u_i = u(e^\alpha_m \otimes 1)$ is the image of $e^\alpha_m \otimes 1$ under the action of u on $R_N \otimes U^+(Witt_{\geq -1})$. In particular if $U^+(Witt_{\geq -1})$ maps some element $v \in V$ to V then it also maps $e^\alpha_i(v)$ to V for any $\alpha \in L$ and $i \in \mathbf{Z}$.*

Proof. If u is of the form L_i for $i \geq -1$ this can be proved as follows. A standard vertex algebra calculation shows that

$$[L_i, e^\alpha_j] = ((i+1)(\alpha^2/2 - 1) - j)e^\alpha_{i+j}.$$

This shows that lemma 5.5 is true when $u \in Witt_{\geq -1}$. If lemma 5.5 is true for two elements u, u' of $U^+(Witt_{\geq -1})$ then it is true for their product. If it is true for some nonzero integral multiple of $u \in U^+(Witt_{\geq -1})$ then it is true for u because V is torsion-free. To finish the proof we observe that the algebra $U^+(Witt_{\geq -1})$ is generated up to torsion by the elements L_m for $m \geq -1$. This proves lemma 5.5.

Lemma 5.6. *Suppose that V is the vertex algebra of the double cover of some even lattice, with the standard action of $Witt_{\geq -1}$ on $V \otimes \mathbf{Q}$ ([B86]). This action extends to an action of $U^+(Witt_{\geq -1})$ on V.*

Proof. The vertex algebra V is generated from the element 1 by the actions of the operators e^α_n for $n \in \mathbf{Z}$ and $\alpha \in L$. Define $F^n(V)$ by defining $F^0(V)$ to be the space spanned by 1, and defining $F^{n+1}(V)$ to be the space spanned by the actions of operators of the form e^α_m on $F^n(V)$. Then V is the union of the spaces $F^n(V)$, so it is sufficient to prove that each space $F^n(V)$ is preserved by the action of $U^+(Witt_{\geq -1})$. We will prove

this by induction on n. For $n = 0$ is is trivial because $L_n(1) = 0$ for $n \geq -1$. If it is true for n, then it follows immediately from lemma 5.5 that it is true for $n + 1$. This proves lemma 5.6.

The algebra $U^+(Witt_{\geq -1})$ can be $\mathbf{Z}$-graded in such a way that L_m has degree m. This follows easily from the fact that we can find graded liftings of the elements L_m. We define $U^+(Witt_{\geq 1})$ to be the subalgebra generated by the coefficients of graded liftings of the elements L_m for $m \geq 1$. It is easy to check that this is a Hopf algebra with a structural basis whose Lie algebra of primitive vectors has a basis consisting of the elements L_m for $m \geq 1$. It is $\mathbf{Z}_{\geq 0}$-graded, with all graded pieces being finite dimensional; in fact the piece of degree $n \in \mathbf{Z}$ has dimension $p(n)$ where p is the partition function.

We let $U(Witt_{\geq 1})$ be the universal enveloping algebra of the Lie algebra $Witt_{\geq 1}$. It is $\mathbf{Z}_{\geq 0}$-graded in the obvious way and is a subalgebra of $U^+(Witt_{\geq 1})$.

In the rest of this section we construct an integral form with a structural basis for the universal enveloping algebra of the Virasoro algebra over $\mathbf{Z}$. This result is not used elsewhere in this paper. We recall that the Virasoro algebra Vir is a central extension of $Witt$ and is spanned by elements L_i for $i \in \mathbf{Z}$ and an element $c/2$ in the center, with

$$[L_m, L_n] = (m - n)L_{m+n} + \binom{m+1}{3}\frac{c}{2}.$$

We identify $Witt_{\geq -1}$ with the subalgebra of Vir spanned by L_m for $m \geq -1$. The Virasoro algebra Vir has an automorphism ω of order 2 defined by $\omega(L_m) = -L_{-m}$, $\omega(c) = -c$, and ω extends to an automorphism of the universal enveloping algebra $U(Vir \otimes \mathbf{Q})$. We define $U^+(Vir)$ to be the subalgebra of $U(Vir \otimes \mathbf{Q})$ generated by $U^+(Witt_{\geq -1})$ and $\omega(U^+(Witt_{\geq -1}))$.

For any even integral lattice L there is a double cover $\hat{L}$, unique up to non-unique isomorphism, such that $e^a e^b = (-1)^{(a,b)} e^b e^a$. We let $V_{\hat{L}}$ be the (integral form of the) vertex algebra of $\hat{L}$. This is L-graded, and has a self dual bilinear form on it (more precisely, each piece of given L-degree α and given eigenvalue under L_0 is finite dimensional and dual to the piece of degree $-\alpha$), and if L is self dual then $V_{\hat{L}}$ has a conformal vector generating an action of the Virasoro algebra.

Theorem 5.7. *The subalgebra $U^+(Vir)$ of $U(Vir \otimes \mathbf{Q})$ has the following properties:*

1. *$U^+(Vir)$ is a $\mathbf{Z}$-Hopf algebra with a structural basis.*
2. *The Lie algebra of primitive elements of $U^+(Vir)$ has a basis consisting of the elements L_n for $n \in \mathbf{Z}$ and the element $c/2$.*
3. *$U^+(Vir)$ maps the vertex algebra (over $\mathbf{Z}$) of any even self dual lattice to itself.*

Proof. We first construct the action on the vertex algebra V of an even self dual lattice. We have an action of $U^+(Witt_{\geq -1})$ on V by lemma 5.6. The vertex algebra of an even self dual lattice is also self dual under its natural bilinear form, so the adjoint of any linear operator on V is also a linear operator on V. The adjoint of L_m is L_{-m}, so the adjoint $U^+(Witt_{\leq 1}) = \omega(U^+(Witt_{\geq -1}))$ also maps V to itself. As these two algebras generate $U^+(Vir)$ this proves that $U^+(Vir)$ acts on V.

Next we find the Lie algebra P of primitive elements of $U^+(Vir)$. The elements L_m for $m \geq -1$ are obviously in $U^+(Vir)$ because they are in $U^+(Witt_{\geq -1})$, and similarly L_m

for $m \leq 1$ is in P. The element $c/2$ is in P because $[L_2, L_{-2}] = 4L_0 + c/2$. So P contains the basis described in theorem 5.7, and we have to prove that P contains no elements other than linear combinations of these. As $U^+(Vir)$ and hence P are both $\mathbf{Z}$-graded with L_m having degree m, it is sufficient to show that the degree m piece of P is spanned by L_m if $m \neq 0$ and by L_0 and $c/2$ if $m = 0$. For $m \neq 0$ this is easy to check as we just map the Virasoro algebra to the Witt algebra and use example 5.3. The case $m = 0$ is harder and we will use the actions on vertex algebras of even self dual lattices L constructed above. If L is such a lattice then c acts on V as multiplication by $\dim(L)$, and L_0 has eigenspaces with eigenvalue any given positive integer (at least if L has positive dimension). So if $xL_0 + yc/2$ is in P for some $x, y \in \mathbf{Q}$ then $xm + yn/2$ is an integer whenever m is a positive integer and n is the dimension of a nonzero even self dual lattice. As we can find such lattices for any positive even integer n, this implies that x and y are both integers. Hence the degree 0 piece of P is spanned by L_0 and $c/2$. This completes the proof that P is spanned by $L_m, m \in \mathbf{Z}$, and $c/2$.

Finally we have to show that $U^+(Vir)$ has a structural basis. We know that all the elements L_m for $m \geq -1$ are liftable in $U^+(Witt_{\geq -1})$, so they are also liftable in $U^+(Vir)$. Similarly the elements L_m for $m \leq 1$ are liftable in $\omega(U^+(Witt_{\geq -1}))$ and therefore in $U^+(Vir)$. It is trivial to check that c is liftable as it acts as multiplication by some integer, so every primitive element of $U^+(Vir)$ is liftable by lemma 2.5 and part 2 of theorem 5.7. Also $U^+(Vir)$ is generated by the coefficients of group-like elements because this is true for $U^+(Witt_{\geq -1})$ and its conjugate under ω. The fact that $U^+(Vir)$ has a structural basis now follows from theorem 2.15. This completes the proof of theorem 5.7.

Example. If L is an even lattice of odd dimension then $L_{-2}(1)$ is not in the vertex algebra of L, because $[L_2, L_{-2}] = 4L_0 + \dim(L)/2$. So part 3 of theorem 5.7 is false without the assumption that L is self dual.

6. The no-ghost theorem over Z.

The no-ghost theorem of Goddard and Thorn [G-T] implies that over the reals, the contravariant form restricted to the degree $\beta \neq 0$ piece of the fake monster Lie algebra $\mathbf{m}$ is positive definite and in particular nonsingular. We give a refinement of this to the integral form of the degree β piece $\mathbf{m}_\beta$, showing that the any prime dividing the discriminant of the quadratic form on $\mathbf{m}_\beta$ also divides the vector β. In particular if β is a primitive vector then $\mathbf{m}_\beta$ is a self dual positive definite integral lattice. I do not know whether or not the discriminant can be divisible by p when β is divisible by p.

If $\lambda = 1^{i_1}2^{i_2}\cdots$ is a partition with i_1 1's, i_2 2's, and so on, then we define $P(\lambda)$ to be the integer $1^{i_1}2^{i_2}\cdots$ and $F(\lambda)$ to be $i_1!i_2!\cdots$ and $|\lambda|$ to be $1i_1 + 2i_2 + \cdots$ and $l(\lambda)$ to be $i_1 + i_2 + \cdots$. We also define $p(n)$ to be the number of partitions of n.

Lemma 6.1. *Suppose that n is an integer. Then*

$$\prod_{|\lambda|=n} P(\lambda) = \prod_{|\lambda|=n} F(\lambda).$$

Proof. We will show that both sides are equal to

$$\prod_{i>0} i^{\Sigma_{j>0} p(n-ij)}.$$

The left hand side $\prod_{|\lambda|=n} P(\lambda)$ is equal to $\prod_{i>0} i^{n(i)}$ where $n(i)$ is the number of times i occurs in some partition of n, counted with multiplicities. This number $n(i)$ is equal to $\sum_j n(i,j)$ where $n(i,j)$ is the number of times that i occurs at least j times in a partition of n. But $n(i,j)$ is equal to $p(n-ij)$ because any partition of n in which i occurs at least j times can be obtained uniquely from a partition of $n-ij$ by adding j copies of i. So the left hand side $\prod_{|\lambda|=n} P(\lambda)$ is equal to $\prod_{i>0} i^{\Sigma_{j>0} p(n-ij)}$.

On the other hand the right hand side $\prod_{|\lambda|=n} F(\lambda)$ is equal to $\prod_{i>0} i^{m(i)}$ where $m(i)$ is the number of times that there is a partition of n with some number occurring at least i times (counting a partition several times if it has more than one number occurring at least i times). But $m(i)$ is equal to $\sum_j n(j,i) = \sum_j p(n-ji)$. This shows that the right hand side $\prod_{|\lambda|=n} F(\lambda)$ is also equal to $\prod_{i>0} i^{\Sigma_{j>0} p(n-ij)}$. This proves lemma 6.1.

Lemma 6.2. *The submodule $U(Witt_{\geq 1})_n$ has index $\prod_{|\lambda|=n} F(\lambda)$ inside $U^+(Witt_{\geq 1})_n$.*

Proof. Choose a graded lifting $1 + a_{i,1}x + a_{i,2}x^2 + \dots$ of L_i for each $i > 0$. Then the elements $a_{1,i_1} a_{2,i_2} \cdots$ for $1i_1 + 2i_2 + \cdots = n$ form a base for $U^+(Witt_{\geq 1})$, and the elements $i_1! a_{1,i_1} i_2! a_{2,i_2} \cdots$ for $1i_1 + 2i_2 + \cdots = n$ form a basis for $U(Witt_{\geq 1})_n$. Therefore the index of $U(Witt_{\geq 1})_n$ in $U^+(Witt_{\geq 1})_n$ is

$$\prod_{i_1+2i_2+\cdots=n} i_1! i_2! \cdots = \prod_{|\lambda|=n} F(\lambda).$$

This proves lemma 6.2.

Lemma 6.3. *Let γ be a norm 0 vector of $II_{25,1}$. Suppose that W is the graded space of all elements generated by the action of the elements $e^{-\gamma} D^{(i)}(e^{\gamma})$ on e^{β}, so that W is acted on by the smooth integral form $U = U^+(Witt_{>0})$. Then the graded dual $W[1/(\beta,\gamma)]^*$ of $W[1/(\beta,\gamma)]$ is a free $U[1/(\beta,\gamma)]$-module on one generator.*

Proof. We define U_n and W_n to be the degree n pieces of U and W. Let w_μ be the basis of elements $\gamma(1)^{j_1} \gamma_2^{j_2} \cdots e^{\beta}$ for $W_n \otimes \mathbf{Q}$ parameterized by partitions $\mu = 1^{j_1} 2^{j_2} \cdots$ of n. The $\mathbf{Z}$ module W_n' spanned by the w_λ's is not W_n but has index $\prod_{|\mu|=n} F(\mu)$ in it. We let the elements $L_\lambda = L_1^{i_1} L_2^{i_2} \cdots$ be the basis for the space $U_n \otimes \mathbf{Q}$ indexed by partitions $\lambda = 1^{i_1} 2^{i_2} \cdots$. The $\mathbf{Z}$ module U_n' spanned by the L_λ's is not U_n but has index $\prod_{|\lambda|=n} F(\lambda)$ in it by lemma 6.2. We define $m_{\lambda,\mu}$ for $|\lambda| = |\mu|$ by $L_\lambda(w_\mu) = m_{\lambda,\mu} e^{\beta}$. We will show that the determinant of the $p(n)$ by $p(n)$ matrix $(m_{\lambda,\mu})$ is

$$\prod_{|\lambda|=n} (\beta,\gamma)^{l(\lambda)} P(\lambda) F(\lambda)$$

where $l(\lambda)$ is the number of elements of the partition λ. We order the partitions by $\lambda > \mu$ if λ is the partition $\lambda_1 + \lambda_2 + \cdots$ with $\lambda_1 \geq \lambda_2 \geq \cdots$, μ is the partition $\mu_1 + \mu_2 + \cdots$ with $\mu_1 \geq \mu_2 \geq \cdots$, and $\lambda_1 = \mu_1, \dots, \lambda_{k-1} = \mu_{k-1}$, $\lambda_k > \mu_k$ for some k. Then the matrix entry $m_{\lambda,\mu}$ is 0 if $\lambda > \mu$, and is equal to

$$P(\lambda) F(\Lambda) (\beta,\gamma)^{l(\lambda)}$$

if $\lambda = \mu$. We can see this by repeatedly using the relation

$$\begin{aligned} &\cdots L_{i_2} L_{i_1} \gamma(j_1)\gamma(j_2)\cdots e^\beta \\ &= \begin{cases} (\text{number of } j\text{'s equal to } j_1)(\beta,\gamma) j_1 \cdots L_{i_2}\gamma(i_2)\cdots e^\beta & \text{if } i_1 = j_1 \\ 0 & \text{if } i_1 > j_1 \end{cases} \end{aligned}$$

for $j_1 \geq j_2 \geq \cdots$, which in turn follows from the identities $[L_i, \gamma(j)] = j\gamma(j-i)$, $L_i(e^\beta) = 0 = \gamma(-i)(e^\beta)$ for $i > 0$, and $\gamma_0 e^\beta = (\beta, \gamma)$ As the matrix $(m_{\lambda,\mu})$ is triangular its determinant is given by the product of the diagonal entries $m_{\lambda,\lambda} = P(\lambda)F(\Lambda)(\beta,\gamma)^{l(\lambda)}$.

Now we work out the index of $U^+(Witt_{>0})(e^{\beta^*})_n$ in W_n^* where $e^{\beta *}$ is the basis element of W_0^* dual to $e^\beta \in W_0$. This index is equal to

$$\frac{(\text{Index of } U_n' \text{ in } W_n')}{(\text{Index of } U_n' \text{ in } U_n)(\text{Index of } W_n' \text{ in } W_n)}$$

The numerator of this expression is equal to the determinant of the matrix $(m_{\lambda,\mu})$, and we calculated this earlier. Substituting in the known values we find the index of $U^+(Witt_{>0})(e^{\beta *})_n$ in W_n^* is

$$\frac{\prod_{|\lambda|=n} P(\lambda)F(\lambda)(\beta,\gamma)^{l(\lambda)}}{(\prod_{|\lambda|=n} F(\lambda))(\prod_{|\mu|=n} F(\mu))}$$

Applying lemma 6.1 we see that this is equal to $\prod_{|\lambda|=n}(\beta,\gamma)^{l(\lambda)}$. This is a unit in $\mathbf{Z}[1/(\beta,\gamma)]$, so over the ring $\mathbf{Z}[1/(\beta,\gamma)]$ the map from $U^+(Witt_{>0})$ to W^* is an isomorphism. This proves lemma 6.3.

Fix a norm 0 vector $\gamma \in L$. We recall that the transverse space if the subspace of elements $v \in V$ such that $L_i(v) = 0$ for $i > 0$, $L_0(v) = v$, and $\gamma(i)(v) = 0$ for $i < 0$. It is easy to check that the transverse space T_β of degree β with $(\beta,\gamma) \neq 0$ is positive definite, and the no-ghost theorem [G-T] works by showing that the natural map from T_β to the space of physical states (modulo null vectors) of degree β is an isomorphism.

Lemma 6.4. *The transverse space T_β of degree $\beta \in II_{1,1}$ has determinant dividing a power of (β,γ).*

Proof. For any vector v of V there is a $U^+(Witt_{>0})$ equivariant map from W^* to V taking 1 to v over $\mathbf{Z}[1/(\beta,\gamma)]$ by lemma 6.3. Dualizing we see that this means there is a vector in $W \otimes V$ of the form $v \otimes 1+$ (terms involving some $\gamma(i)$) which is fixed by $U^+(Witt_{>0})$ and hence also fixed by $Witt_{>0}$. This gives a map from V to the transverse space which is an isomorphism over $\mathbf{Z}[1/(\beta,\gamma)]$. This isomorphism preserves the bilinear form because γ has norm 0 so all the terms involving $\gamma(i)$ have zero inner product with all other terms. Hence the transverse space is self dual over $\mathbf{Z}[1/(\gamma,\beta)]$. This proves lemma 6.4.

The next theorem is an extension of the no-ghost theorem from rational vector spaces to modules over $\mathbf{Z}$. Recall that the usual no-ghost theorem [G-T] says that if $\alpha \in II_{25,1}$ is nonzero then the space of physical states (over $\mathbf{Q}$) of degree α is positive definite and spanned by the transverse space, which has dimension $p_{24}(1 - \alpha^2/2)$. This describes the space of physical states as a rational vector space, but it also has a natural lattice inside it and we can ask about the structure of this lattice.

Theorem 6.5. (The no-ghost theorem over Z). *Suppose α is a nonzero vector of $II_{25,1}$ which is n times a primitive vector. Then the discriminant of the space of physical states of degree α divides a power of n. In particular the space of physical states is a self dual lattice if β is primitive.*

Proof. For every prime p coprime to n we can find a norm zero vector γ with (γ, β) coprime to p, because the norm 0 vectors of Leech type span $II_{25,1}$. By lemma 6.4 this implies that the discriminant of the space of physical states of degree β is coprime to p. Hence this discriminant divides a power of n. This proves theorem 6.5.

The proof of this theorem implies that all spaces of physical states of primitive vectors of the same norm are isomorphic over all p-adic fields, in other words in the same genus. It is certainly not always true that they are isomorphic over the integers. For example, for a norm 0 vector the space of physical states is isomorphic to the corresponding Niemeier lattice, and not all Niemeier lattices are isomorphic.

7. An application to modular moonshine.

In this section $\mathbf{m}$ will stand for the fake monster Lie algebra over $\mathbf{Z}[1/2]$ (see [B-R], [B98]) rather than the fake monster Lie algebra.

We write $\mathbf{Z}_p$ for the ring of p-adic numbers. The paper [B98] showed that Ryba's modular moonshine conjectures [R] for primes $p \geq 13$ were true provided the following assumption was true:

Assumption. *If $m < p$ then the degree (m, n) piece of $\mathbf{m} \otimes \mathbf{Z}_p$ is self dual under the natural bilinear form and isomorphic to $V_{mn} \otimes \mathbf{Z}_p$ as a $\mathbf{Z}_p$ module acted on by the monster.*

In this section we will prove this assumption, thus completing the proof of the modular moonshine conjectures for $p \geq 13$. In fact we will prove the following slightly stronger theorem:

Theorem 7.1. *If m and n are not both divisible by p then the degree (m, n) piece of $\mathbf{m} \otimes \mathbf{Z}_p$ is self dual under the natural bilinear form and isomorphic to $V_{mn} \otimes \mathbf{Z}_p$ as a $\mathbf{Z}_p$ module acted on by the monster.*

Proof. If one of m or n is not divisible by p then we can find a norm 0 vector γ with (γ, β) coprime to p. Theorem 7.1 then follows from the integral no-ghost theorem 6.5, because the discriminant of $\mathbf{m}_{(m,n)}$ is not divisible by p and is therefore a unit in $\mathbf{Z}_p$. Note that the isomorphism given by theorem 6.5 preserves the action of the monster. This proves theorem 7.1.

8. Open problems.

1 The integral form for the fake monster Lie algebra constructed in section 4 is probably not the best possible one. Is there an integral form which not only has a structural basis but also has the property that the bilinear form on the primitive elements is self dual over $\mathbf{Z}$?

2 Can the integral no-ghost theorem 6.5 be extended to show that the root spaces of non-primitive vectors are self dual, and in particular is the bilinear form on $\mathbf{m}$ self dual? We have shown in theorem 6.5 that it is self dual on the root spaces of roots

that are either primitive or of norm 0. This question may be irrelevant, because it is possible to construct an integral form of the fake monster Lie algebra with a self dual bilinear form using the BRST complex. Over the complex numbers this construction is described in [F-G-Z] and [L-Z]. The construction can be done over the integers, and the BRST complex turns out to be a vertex superalgebra isomorphic to the vertex superalgebra of the odd self dual lattice $I_{26,1}$. The fake monster Lie algebra is given by the homology groups of the self adjoint BRST operator Q modulo torsion, and the homology modulo torsion of a complex with a self dual bilinear form is automatically self dual under the induced form. However it is not clear that this self dual integral form can be extended to a smooth Hopf algebra over $\mathbf{Z}$.

3 Which elements of $\mathbf{m}_\alpha$ are liftable, and in particular are all elements liftable?

4 We can also ask all the questions above about the monster Lie algebra rather than the fake monster Lie algebra. In this case much less is known than for the fake monster Lie algebra. In particular I do not even know of an integral self dual form of the monster vertex algebra or the monster Lie algebra (but see [B-R] for some speculation about this).

5 The arguments in section 4 of [B98] proving the modular moonshine conjectures for $p \geq 13$ are rather messy and computational, mainly because of the lack of a good description of what happens at the root spaces of roots α divisible by p. Is is possible to clean up this argument using the (conjectured) existence of a nice Hopf algebra for the monster Lie algebra?

References.

[A] E. Abe, Hopf algebras. Cambridge Tracts in Mathematics, 74. Cambridge University Press, Cambridge-New York, 1980. xii+284 pp. ISBN: 0-521-22240-0

[B86] R. E. Borcherds, Vertex algebras, Kac-Moody algebras, and the monster. Proc. Natl. Acad. Sci. USA. Vol. 83 (1986) 3068–3071.

[B90] R. E. Borcherds, The monster Lie algebra, Adv. Math. Vol. 83, No. 1, Sept. 1990.

[B98] R. E. Borcherds, Modular Moonshine III, to appear in Duke Math Journal.

[B-R] R. E. Borcherds, A. J. E. Ryba, Modular Moonshine II, Duke Math Journal Vol. 83 No. 2, 435-459, 1996.

[D] J. Dieudonné, Groupes de Lie et hyperalgèbres de Lie sur un corps de caractéristique $p > 0$ (V), Bull. Soc. Math. France 84 1956 207–239 reprinted in J. Dieudonné, "Choix d'œuvres mathématiques. Tome II." Hermann, Paris, 1981, ISBN: 2-7056-5923-4, p. 600–632.

[D72] E. J. Ditters, Curves and formal (co)groups, Inv. Math. 17 (1972) 1-20.

[F-G-Z] I. Frenkel, H. Garland, G. J. Zuckerman, Semi-infinite cohomology and string theory. Proc. Nat. Acad. Sci. U.S.A. 83 (1986), no. 22, 8442–8446.

[G-T] P. Goddard and C. B. Thorn, Compatibility of the dual Pomeron with unitarity and the absence of ghosts in the dual resonance model, Phys. Lett., B 40, No. 2 (1972), 235-238.

[H] M. Hazewinkel, "Formal groups and applications". Pure and Applied Mathematics, 78. Academic Press, Inc. [Harcourt Brace Jovanovich, Publishers], New York-London, 1978. ISBN: 0-12-335150-2

[K] B. Kostant, Groups over Z. 1966 Algebraic Groups and Discontinuous Subgroups (Proc. Symposium. Pure Math., Boulder, Colorado., 1965) pp. 90–98 Amer. Math. Soc., Providence, R.I.

[L-Z] B. Lian, G. J. Zuckerman, Moonshine cohomology. Moonshine and vertex operator algebra (Kyoto, 1994). Sūrikaisekikenkyūsho Kōkyūroku No. 904 (1995), 87–115.

[M] I. G. Macdonald, Symmetric functions and Hall polynomials. Second edition. Oxford Mathematical Monographs. Oxford Science Publications. The Clarendon Press, Oxford University Press, New York, 1995. x+475 pp. ISBN: 0-19-853489-2

[R] A. J. E. Ryba, Modular Moonshine?, In "Moonshine, the Monster, and related topics", edited by Chongying Dong and Geoffrey Mason. Contemporary Mathematics, 193. American Mathematical Society, Providence, RI, 1996. 307-336.

[Sh] P. B. Shay, An obstruction theory for smooth formal group structure, Preprint, Hunter college, CUNY.

arXiv:math/9806141v2 [math.GR] 26 Jul 1998

Coxeter groups, Lorentzian lattices, and K3 surfaces. *25 June, 27 July 1998*

Richard E. Borcherds, *
D.P.M.M.S., 16 Mill Lane, Cambridge, CB2 1SB, England.
e-mail: reb@dpmms.cam.ac.uk
www home page www.dpmms.cam.ac.uk/~reb

Contents.

1. Introduction.
2. Notation and statement of main theorem.
3. Proof of main theorem.
4. The structure of Γ_Ω.
5. Examples.

1. Introduction.

The main result of this paper describes the normalizer $N_{W_\Pi}(W_J)$ of a finite parabolic subgroup W_J of a (possibly infinite) Coxeter group W_Π. More generally we describe $N_{W_\Pi.\Gamma_\Pi}(W_J)$ where Γ_Π is a group of diagram automorphisms of the Coxeter diagram Π of W_Π. By taking Π to be Conway's Coxeter diagram of the reflection group of $II_{1,25}$ we compute the automorphism groups of some Lorentzian lattices and K3 surfaces.

In the case when W_Π is a finite Coxeter group (and $\Gamma_\Pi = 1$) the normalizer of W_J has been described by Howlett [H]. His result states that $N_{W_\Pi}(W_J)$ is a split extension $W_J.W'_J$, where $W'_J = W_\Omega.\Gamma_\Omega$ is in turn a split extension with W_Ω a Coxeter group and Γ_Ω a more mysterious group acting on Ω. Howlett showed by case by case analysis that if Π is connected (and W_Π is finite) then Γ_Ω is an elementary abelian 2-group and is a subgroup of $Aut(J)$. When W_Π is infinite the normalizer $N_{W_\Pi}(W_J)$ has a similar structure, except that the group Γ_Ω can be more complicated. Although there is still a canonical map from Γ_Ω to $Aut(J)$, the kernel can be non trivial, though it has finite cohomological dimension. The kernel is trivial in the case of finite W_Π considered by Howlett because any finite group of finite cohomological dimension must be trivial. For example, the case when $J = A_1$, $\Gamma_\Pi = 1$ has been done by Brink [Br], who showed that Γ_Ω is a free group (and therefore has cohomological dimension at most 1). We will extend Brink's result to Coxeter diagrams of arbitrary finite reflection groups. More precisely we construct a category Q_4 using Π and J and prove that the classifying space of this category is a classifying space of the group Γ_Ω. The main point about this category Q_4 is that it is often finite and can often be written down explicitly, in which case we can easily read off a presentation of Γ_Π. For example, if $J = A_1$ we show that the classifying space of this category Q_4 is 1-dimensional, so its fundamental group is free and we recover Brink's result. After writing this paper I discovered that Brink and Howlett had previously announced a related description of the normalizer of a parabolic subgroup of a Coxeter group; see [B-H], and see example 2.8 for the relation between their result and theorem 2.7.

For later applications we need some generalizations as follows. First of all, instead of calculating the normalizer of W_J in a Coxeter group W_Π, we calculate the normalizer in an

* Supported by a Royal Society professorship and an NSF grant.

extension $W_\Pi.\Gamma_\Pi$, where Γ_Π is a group of diagram automorphisms. Secondly, we sometimes want to compute not the full normalizer, but a subgroup with image contained in some subgroup Γ_J of $Aut(J)$. Thirdly, we sometimes want to vary the choice of the Coxeter group W_Ω, which we do by varying a certain normal subgroup R of Γ_J. For example, in calculating the automorphism groups of $K3$ surfaces we take W_Ω to be generated by reflections of norm -2 vectors rather than by all reflections, so we take $R = 1$.

Section 2 contains a statement of the main result (theorem 2.7) describing a classifying category for the group Γ_Ω, and section 3 contains the proof of this result. Section 4 contains some more information about the structure of Γ_Ω. In section 5 we give some applications of theorem 2.7, and in particular show how to describe the automorphism groups of some Lorentzian lattices by embedding them in $II_{1,25}$ and using the description of $Aut(II_{1,25})$ in [C]. The idea of studying Lorentzian lattices by embedding them as orthogonal complements of root lattice in $II_{1,25}$ comes from Conway and Sloane ([C-S]).

Work of I. Piatetski-Shapiro and I. R. Shafarevich [P-S] shows that there is a map from the automorphism group of a K3 surface to the group of automorphisms of its Picard lattice modulo the group generated by reflections of norm -2 vectors which has finite kernel and co-finite image, so in practice if we want to describe the automorphisms of a K3 surface the main step is to calculate the automorphism group of its Picard lattice. Kondo showed in [K] that the automorphism groups of some K3 surfaces could be studied by embedding their Picard lattice as the orthogonal complement of a root lattice in $II_{1,25}$. We use Kondo's idea to describe the automorphism groups of some K3 surfaces in terms of combinatorics of the Leech lattice. In particular we reprove some results of Vinberg [V] on the "most algebraic" K3 surfaces and extend them to the "next most algebraic" K3 surface. Kondo showed in [K] that the automorphism group of the Kummer surface of a generic genus 2 Jacobian was generated by the classically known automorphisms together with some new automorphisms found by Keum [Ke], and we show how to use Kondo's results to describe the structure of this group. Kondo and Keum [K-K] have recently proved similar results for some Kummer surfaces associated to the products of two elliptic curves.

Kondo recently found another mysterious connection between automorphism groups of K3 surfaces and Niemeier lattices [K98], and used this to give a short proof of Mukai's classification [Mu] of the finite groups that act on K3 surfaces.

I would like to thank I. Cherednik, I. Grojnowski, R. B. Howlett, J. M. E. Hyland, S. Kondo, U. Ray, and G. Segal for their help.

2. Notation and statement of main theorem.

This section states the main result (theorem 2.7) describing normalizers of parabolic subgroups of Coxeter groups.

We recall some basic definitions about Coxeter systems. For more about them see [Hi] or [Bo]. A pair (W, S) is called a **Coxeter system** if W is a group with a subset S such that W has the presentation

$$\langle s : s \in S | (ss')^{m_{ss'}} = 1 \text{ when } m_{ss'} < \infty \rangle$$

where $m_{ss'} \in \{1, 2, 3, \dots, \infty\}$ is the order of ss', and $m_{ss'} = 1$ if and only if $s = s'$. A diagram automorphism of S is an automorphism of the set S that extends to an automor-

phism of the group W, and $Aut(S)$ means the group of diagram automorphisms of S. We say that (W, S) is **irreducible** if S is not a union of two disjoint commuting subsets. The number of elements of S is called its **rank**. The Coxeter system is called **spherical** if W has finite order. The irreducible spherical Coxeter diagrams are A_n $(n \geq 1)$, $B_n = C_n$ $(n \geq 2)$, D_n $(n \geq 4)$, E_6, E_7, E_8, F_4, $G_2^{(n)} = I_2(n)$ $(n \geq 5)$, H_3, and H_4. It is also sometimes useful to define the Coxeter diagrams $B_1 = C_1 = A_1$, $D_3 = A_3$, $D_2 = A_1^2$, $E_5 = D_5$, $E_4 = A_4$, $E_3 = A_2A_1$, $G_2^{(4)} = B_2 = C_2$, $G_2^{(3)} = A_2$, $G_2^{(2)} = A_1^2$.

If (W_Π, Π) is a Coxeter system then we write V_Π for the (possibly infinite dimensional) real vector space with a basis of elements e_s for $s \in \Pi$, and put a symmetric bilinear form on V_Π by defining

$$(e_s, e_{s'}) = 2\cos(\pi/m_{ss'}).$$

Note that we normalize the roots e_s so that they have norm $(e_s, e_s) = -2$ rather than 1; this is done to be consistent with the usual conventions in algebraic geometry.

The Coxeter group W_Π acts on V_Π with the element $s \in \Pi \subseteq W_\Pi$ acting as the reflection $v \mapsto v + (v, e_s)e_s$ in the hyperplane $e_s^\perp$. Any subgroup Γ_Π of $Aut(\Pi)$ acts on V_Π by permutations of the elements e_s, so we get an action of $W.\Gamma_\Pi$ on V_Π, and hence on the dual space V_Π^*. We write Δ^+ for the set of positive roots of W. We define the fundamental chamber $C_\Pi \subseteq V_\Pi^*$ of W_Π by

$$C_\Pi = \{x \in V_\Pi^* | x(r) \geq 0 \text{ for all } r \in \Pi \text{ and } x(r) > 0 \text{ for almost all } r \in \Delta^+\}.$$

(Recall that "almost all" means "all but a finite number of".) A theorem due independently to Tits and Vinberg states that no two distinct points of C_Π are conjugate under W, and the subgroup of W fixing all points of some subset A of W is generated by the reflections in the faces of W containing A. In particular W acts simply transitively on the conjugates of C_Π. The union $W_\Pi(C_\Pi)$ of all conjugates of C_Π under W_Π is given by

$$W_\Pi(C_\Pi) = \{x \in V_\Pi^* | x(r) > 0 \text{ for almost all } r \in \Delta^+\}.$$

In particular $W_\Pi(C_\Pi)$ is convex and closed under multiplication by positive real numbers. If $x \in W_\Pi(C_\Pi)$ then the set of roots vanishing on x is a finite root system.

Note that $W_\Pi(C_\Pi)$ is usually slightly smaller than the Tits cone, which is defined in the same way except that we omit the condition that $x(e_s) > 0$ for all but a finite number of s in the definition of the fundamental domain. The Tits cone can be thought of as obtained from $W_\Pi(C_\Pi)$ by "adding some boundary components". The reason for using $W_\Pi(C_\Pi)$ rather than the Tits cone is that the cone $W_\Pi(C_\Pi)$ has the property that the subgroup of W_Π fixing any vector of it is finite.

We fix a spherical subset J of Π. In particular we get a spherical Coxeter system (W_J, J). Suppose K is an isometry of Coxeter diagrams from J into Π. We write W_K for the finite reflection group generated by $K(J)$. There is a natural homomorphism $p : N_{W_\Pi.\Gamma_\Pi}(W_K) \mapsto Aut(J)$. We let $N_{W_\Pi.\Gamma_\Pi}(W_K; \Gamma_J)$ be the subgroup of elements whose image is in a subgroup Γ_J of $Aut(J)$. We are interested in describing the group $N_{W_\Pi.\Gamma_\Pi}(W_J; \Gamma_J)$. Most of the time we take $\Gamma_J = Aut(J)$ in which case $N_{W_\Pi.\Gamma_\Pi}(W_J; \Gamma_J) = N_{W_\Pi.\Gamma_\Pi}(W_J)$, but it is occasionally useful to use other values of Γ_J; see example 5.7 and theorem 4.1.

We define W_K' to be the subgroup of $N_{W_\Pi.\Gamma_\Pi}(W_K; \Gamma_J)$ mapping $K(J)$ to itself.

Lemma 2.1. $N_{W_\Pi.\Gamma_\Pi}(W_K;\Gamma_J) = W_K.W'_K$.

Proof. This follows immediately from the fact that W_K acts simply transitively on the Weyl chambers of W_K, and W'_K is the subgroup of $N_{W_\Pi.\Gamma_\Pi}(W_K;\Gamma_J)$ fixing a Weyl chamber of W_K. This proves lemma 2.1.

The group W'_K acts on the subspace $V_\Pi^{*W_K}$ of V_Π^* of all vectors fixed by W_K. We now construct a reflection group W_{Ω_K} acting on $V_\Pi^{*W_K}$. We choose a normal subgroup R of Γ_J. (The subgroup R is used to control the reflection group W_Ω defined below. Often we want W_Ω to be as large as possible and we take $R = \Gamma_J$, but sometimes we want to take a smaller W_Ω; see examples 5.3, 5.4 and 5.5.) We define the group W_{Ω_K} to be the subgroup of $W_\Pi \cap W'_K$ generated by elements $w \in W_\Pi \cap W'_K$ such that w acts on $V_\Pi^{*W_K}$ as a reflection and acts on J as an element of R. We define Ω_K to be the Coxeter diagram of W_{Ω_K} and C_{Ω_K} to be its fundamental chamber. If K is the identity map from J to $J \subseteq \Pi$ then we write W_Ω, C_Ω, and Ω instead of W_{Ω_K}, C_{Ω_K}, and Ω_K. Note that W_{Ω_K} is obviously contained in the inverse image of R in $W_\Pi \cap W'_K$, but can be much smaller; see for example the discussion of D_4 in example 5.7.

We define the group Γ_{Ω_K} to be the subgroup of W'_K of elements w with $w(C_{\Omega_K}) = C_{\Omega_K}$.

Lemma 2.2. *The group W'_K is a semidirect product $W'_K = W_{\Omega_K}.\Gamma_{\Omega_K}$.*

Proof. We first show that the group W_{Ω_K} acts faithfully on $V_\Pi^{*W_K}$. More generally we will show that if $w \in W_\Pi \cap W'_K$ acts trivially on $V_\Pi^{*W_K}$, then $w = 1$. To see this we observe that w fixes the point $x \in C_\Pi \cap V_\Pi^{*W_K}$ such that $x(e_s) = 0$ if $s \in K(J)$ and $x(e_s) = 1$ if $s \notin K(J)$. Therefore w is in the subgroup W_K of W_Π generated by the simple reflections of W_Π fixing x. On the other hand w maps $K(J)$ into itself as $w \in W'_K$. This implies that $w = 1$ because 1 is the only element of W_K mapping $K(J)$ into itself. This proves that the group W_{Ω_K} acts faithfully on $V_\Pi^{*W_K}$.

Lemma 2.2 now follows from the fact that W_{Ω_K} acts simply transitively on the conjugates of C_{Ω_K} under W'_K, and Γ_{Ω_K} is the stabilizer of C_{Ω_K}. This proves lemma 2.2.

Warning: the group Γ_{Ω_K} need not act faithfully on $V_\Pi^{*W_K}$ (though it does act faithfully on $V_\Pi^{*W_K} \times J$).

We define a **classifying category** of a group Γ to be a category whose geometric realization is a classifying space for Γ. (Recall from [Q] that the geometric realization of a category is a space with a 0-cell for each object and an n-cell for each sequence of n composable morphisms if $n > 0$.) For example, the category with one object whose morphisms are the elements of Γ (with composition given by group multiplication) is a classifying category for Γ.

We have more or less reduced the problem of describing $N_{W_\Pi.\Gamma_\Pi}(W_J;\Gamma_J)$ to that of describing Γ_{Ω_K}. (The Coxeter diagram Ω of W_Ω can be described once we know Γ_Ω.) The main theorem of this paper describes Γ_{Ω_K} by giving an explicit classifying category for it. To define this category we need some more definitions.

Suppose that S is the Coxeter diagram of a finite reflection group G of a finite dimensional vector space with no vectors fixed by G. Fix a Weyl chamber C of G, so that the walls of C correspond to the points of S. Then there is a unique element σ_S of G taking C to $-C$ called the opposition involution. The involution $-\sigma_S$ acts on the the Coxeter

diagram S, and its action on S does not depend on the choice of C. This action can be described as follows. The points of the Coxeter diagram correspond to the simple roots of C. This set of roots is the same as the set of simple roots of $\sigma_S(C) = -C$ multiplied by -1. Hence $-\sigma_S$ acts on this set of simple roots, in other words on the Coxeter diagram S. The involution $-\sigma_S$ of S can be described explicitly as follows. On diagrams of type A_1, $B_n = C_n$, D_{2n}, E_7, E_8, F_4, $G_2^{(2n)}$, H_3, H_4 for any $n \geq 1$ the involution $-\sigma_S$ is the trivial automorphism of S, while for diagrams of types A_{n+1}, D_{2n+1}, E_6, $G_2^{(2n+1)}$ for $n \geq 1$ the involution $-\sigma_S$ is the unique nontrivial automorphism of the Coxeter diagram S. Finally if the diagram S is a union of connected components then $-\sigma_S$ acts on each connected component as described above.

Suppose that J and S are Coxeter diagrams. Suppose that K and K' are two isometries from J into S. We define K and K' to be **adjacent** if there is a point s of S not in $K(J)$ such that $K(J) \cup s$ is spherical and $\sigma_{K(J)\cup s}\sigma_{K(J)}$ takes K to K'. If K is adjacent to K' then K' is adjacent to K. We define two isometries K, K' of J into S to be **associate** if there is a sequence of isometries $K = K_1, K_2, \ldots, K_n = K'$ from J into S such that K_i and K_{i+1} are adjacent for $1 \leq i < n$. We write $\overline{K}$ for the equivalence class of all isometries from J into S that are associate to K.

Remark. The term associate is closely related to the same term in algebraic group theory, where two parabolic subgroups are called associate if their Levi factors are conjugate. There is a small difference because a standard parabolic subgroup corresponds to a subdiagram of the Dynkin diagram, while we work with *labeled* subdiagrams of a Coxeter diagram.

Suppose that S is a spherical Coxeter diagram. Recall that R is a normal subgroup of $\Gamma_J \subseteq Aut(J)$. We define an isometry $K : J \mapsto S$ to be **R-reflective** if K is adjacent to $r(K)$ for some $r \in R$. We define the equivalence class $\overline{K}$ to be **R-reflective** if at least one of its elements is R-reflective. (The reason for the name "R-reflective" appears in lemmas 3.8 and 3.9.)

Example 2.3. Suppose J is A_1 and S is A_3, with the isometries from J into S labeled as K_1, K_2, K_3 in the obvious way. Then the isometry K_1 is adjacent to K_2 as $\sigma_{K_1(J)\cup K_2(J)}\sigma_{K_1}$ takes K_1 to K_2. Similarly K_2 is adjacent to K_3, and K_1 is associate to K_3 but not adjacent to K_3. The isometries K_1 and K_3 are adjacent to themselves and K_2 is not adjacent to itself, so the isometry K_2 is not R-reflective but the isometries K_1 and K_3 are. So K_2 is not R-reflective but the equivalence class $\overline{K_2}$ is.

Example 2.4. Suppose K is D_5 and S is D_6. Then there are exactly two isometries $K_1, K_2 : J \mapsto S$ of S, which are adjacent to each other but not to themselves. These two isometries are exchanged by the nontrivial automorphism of D_5. Hence if R contains just $1 \in Aut(D_5)$ then K_1, K_2, and the equivalence class $\overline{K_1} = \{K_1, K_2\}$ are not R-reflective, but if R is the whole of $Aut(D_5) = Z/2Z$ then all of them are R-reflective.

Example 2.5 Suppose K is A_3 and S is D_5. Then there are 8 isometries $K : J \mapsto S$. These form two equivalence classes under the relation of being associate, one of size 2 and one of size 6. This shows that two isometries from a connected diagram J into S need not be associate to conjugates of each other under $Aut(J)$.

Example 2.6 Suppose J is A_2. If S is A_n for $n \geq 2$ then there are two equivalence classes of isometries $K : J \mapsto S$, which are exchanged by the nontrivial automorphism of

A_2. However if S is D_n ($n \geq 4$), E_6, E_7, or E_8 then there is only one equivalence class, as the A_2 can be reversed by doing a "three point turn" around the point of valence 3 in S.

We define a poset P_3^+ as follows. The objects of P_3^+ are pairs $(S, \overline{K})$ consisting of a spherical subdiagram S of Π and an equivalence class $\overline{K}$ of isometries from J into S. We define the partial order on P_3^+ by putting $(S, \overline{K}) \leq (S', \overline{K'})$ if $S \subseteq S'$ and $\overline{K} \subseteq \overline{K'}$. We define P_3 to be the sub-poset of P_3^+ of elements $(S, \overline{K})$ such that $\overline{K}$ is not R-reflective.

Note that the condition that $\overline{K}$ is not R-reflective is quite restrictive and implies that S is usually not much larger than J and in any case has at most twice the rank of J. In particular S cannot contain any root orthogonal to $K(J)$ as this implies that K is R-reflective for any R.

Suppose that P is a poset acted on by a group G. We define the **homotopy quotient** of P by G to be the following category Q. The objects of Q are the elements of P. The morphisms from $p_1 \in P$ to $p_2 \in P$ correspond to the group elements $g \in G$ such that $g(p_1) \leq p_2$, and composition of morphisms is given by multiplication of group elements. If G is trivial this is the usual category associated to the poset P, and if P has just one point this is the usual category with one object associated to the group G. If we take a full set of representatives of the orbits of G on P, then the full sub category of Q with these objects is a skeleton of the category Q.

The poset P_3 is acted on by the group $\Gamma_J \times \Gamma_\Pi \subseteq Aut(J) \times Aut(\Pi)$. We define Q_3 to be the homotopy quotient of P_3 by $\Gamma_J \times \Gamma_\Pi$, and we construct the category Q_4 as a skeleton of Q_3 as above. In other words the objects of Q_4 are a complete set of representatives for the orbits of $\Gamma_J \times \Gamma_\Pi$ on the elements of P_3 and the morphisms from $(S, \overline{K})$ to $(S', \overline{K'})$ correspond to group elements $\gamma \in \Gamma_J \times \Gamma_\Pi$ such that $\gamma((S, \overline{K})) \leq (S', \overline{K'})$.

The main result of this paper is the following description of the classifying space of Γ_Ω.

Theorem 2.7. *Suppose we are given the following objects.*

(W_Π, Π) *A Coxeter system.*
Γ_Π *A subgroup of* $Aut(\Pi)$.
(W_J, J) *A spherical Coxeter system with* $J \subseteq \Pi$.
Γ_J *A subgroup of* $Aut(J)$.
R *A normal subgroup of* Γ_J.

Define W_Ω, Γ_Ω, *and* Q_4 *as above, so that the group* $N_{W_\Pi.\Gamma_\Pi}(W_J; \Gamma_J)$ *has the structure*

$$W_J.W_\Omega.\Gamma_\Omega$$

where W_Ω *is a Coxeter group. Then the component of* Q_4 *containing the object* $(J, \overline{id_J})$ *is a classifying category for the group* Γ_Ω.

Theorem 2.7 gives a presentation of the group Γ_Ω because Γ_Ω is the fundamental group of the category Q_4 with respect to the basepoint $(J, \overline{id_J})$, and it is easy to write down a presentation of the fundamental group of any connected category Q as follows. Choose a spanning tree T for the underlying 1-complex of Q (which has a point for each object of Q and a 1-cell for each morphism). Then the fundamental group Γ_Ω of Q has a presentation as follows. The group Γ_Ω has a generator $\overline{g}$ for each morphism g of Q. The relations are $\overline{gh} = \overline{g}\overline{h}$ whenever gh is defined, and $g = 1$ for g in the spanning tree.

Example 2.8. Suppose we put $R = \Gamma_J = Aut(J)$ and $\Gamma_\Pi = 1$. Then we see from theorem 2.7 that $N_{W_\Pi}(W_J) = W_J.W_\Omega.\Gamma_\Omega$, where Γ_Ω is the fundamental group of the component of Q_4 corresponding to J. So in particular theorem 2.7 describes normalizers of finite parabolic subgroups of Coxeter groups. More generally, Brink and Howlett [B-H] have described a presentation for normalizers of possibly infinite parabolic subgroups of Coxeter groups. It is not trivial to see that the presentation given by [B-H] is equivalent to the one given by theorem 2.7 (though of course this follows from the fact that they are both presentations of the same group). Howlett pointed out to me that their result only requires considering subdiagrams S of Π whose rank is at most $2 + rank(J)$ to get the relations of Γ_Ω, and of rank at most $1 + rank(J)$ to get the generators of Γ_Ω. It seems possible that a similar simplification could be made to theorem 2.7 if all that is required is a presentation rather than a classifying space. Perhaps the natural map from $\pi_i(Q_4^j)$ to $\pi_i(Q_4)$ is an isomorphism for $i < j$ and an epimorphism for $i = j$, where Q_4^j is the full sub-category of Q_4 whose objects are the elements $(S, \overline{K})$ such that $rank(S) \leq rank(J) + j$. If so, the map from $\pi_1(Q_4^2)$ to $\pi_1(Q_4)$ would be an isomorphism, so this would give a closer connection to the presentation of Brink and Howlett. Their result also suggests that theorem 2.7 could be generalized by allowing W_J to be infinite and modifying the definition of Q_4 to allow subdiagrams S such that W_J has finite index in W_S.

3. Proof of main theorem.

This section gives the proof of the theorem 2.7. The idea of the proof is to construct categories and functors according to the following diagram.

$$Q_1 \longleftarrow Q_2 \longrightarrow Q_3 \longleftarrow Q_4$$

It is easy to show that a component of Q_1 is a classifying category for Γ_Ω. We also show that the functors between the categories are all homotopy equivalences, so a component of Q_4 is a classifying category for Γ_Ω, which is what we wanted to prove.

We define an isometry from J into the roots of W_Π to be **primitive** if it is conjugate under W_Π to an isometry from J into Π. An example of a non-primitive isometry is an isometry from A_1^4 into the roots of D_4.

We define a category Q_1 as follows. We define the poset P_1 to be the poset of pairs (C, K) where K is a primitive isometry from J into the (possibly non-simple) roots of W_Π, and C is a Weyl chamber of the reflection group W_{Ω_K} of $V_\Pi^{*W_K}$. The partial order on P_1 is the trivial one with $(C_1, K_1) \leq (C_2, K_2)$ if and only if $(C_1, K_1) = (C_2, K_2)$. The objects of P_1 are acted on in the natural way by the group $W_\Pi.\Gamma_\Pi$ via its action on V_Π, and by the group Γ_J via its action on J. We define the category Q_1 to be the homotopy quotient of P_1 by the group $\Gamma_J \times W_\Pi.\Gamma_\Pi$.

Lemma 3.1. *The component of Q_1 containing the object $id_J : J \mapsto J \subseteq \Pi$ is a classifying category for the group Γ_Ω.*

Proof. This follows immediately from the fact that Q_1 is a groupoid such that the automorphism group of the object K is the group Γ_{Ω_K}. This proves lemma 3.1.

Let C_Π be the Weyl chamber of W_Π defined in section 2. By a face of C_Π we mean a nonempty intersection of C_Π with some of the hyperplanes bounding C_Π. The faces of

C_Π of codimension n correspond to the spherical subdiagrams of Π of rank n. We define a Π-cell to be a conjugate of a face of C_Π under W_Π. The cone X is the union of all Π-cells, and the intersection of two Π-cells is either empty or another Π-cell.

We define a category Q_2 and posets P_2, P_2^+ as follows. The objects of the poset P_2^+ are the pairs (D, K) where K is a primitive isometry from J into the roots of W_Π, and D is a Π-cell contained in $V_\Pi^{*W_K}$. We define the partial order on P_2^+ by saying $(D_1, K_1) \leq (D_2, K_2)$ if $D_1 \subseteq D_2$ and $K_1 = K_2$. We define P_2 to be the sub-poset of P_2^+ of elements (D, K) such that D is not contained in a reflection hyperplane of W_{Ω_K}. The category Q_2 is defined to be the homotopy quotient of P_2 by $\Gamma_J \times W_\Pi.\Gamma_\Pi$.

Lemma 3.2. *Suppose G is a group, P_1 and P_2 are G-posets, and f is a morphism of G-posets from P_2 to P_1. Also suppose that for any $Y \in P_1$ the poset $f^{-1}(Y)$ is contractible (in other words the corresponding simplicial complex is contractible). Then the functor induced by f between the homotopy quotient categories Q_2, Q_1 of the posets P_2 and P_1 by the group G is a homotopy equivalence.*

Proof. If f is a functor from a category Q_2 to a category Q_1 and Y is an object of Q_1 then we write $f^{-1}(Y)$ for the fiber of f over Y, in other words the sub category of Q_2 whose morphisms are those mapped to the identity of Y by f. We write $Y\backslash f$ for the category consisting of pairs (X, v) with $v : Y \mapsto f(X)$, where a morphism from (X, v) to (X', v') is a morphism $w : X \mapsto X'$ such that $f(w)v = v'$. Then a result due to Quillen (the corollary to theorem A on page 9 of [Q]) states that f is a homotopy equivalence provided that for all Y in Q_1 the poset $f^{-1}(Y)$ is contractible and the functor from $f^{-1}(Y)$ to $Y\backslash f$ taking X to (X, id_Y) has a right adjoint. (Here id_Y is the identity morphism of Y.)

We will use Quillen's result to show that f is a homotopy equivalence. For any object Y of P_1 the category $f^{-1}(Y)$ is just the category of the poset $f^{-1}(Y)$, which is contractible by assumption. So it only remains to check the condition about the existence of a right adjoint from $Y\backslash f$ to $f^{-1}(Y)$. The category $Y\backslash f$ has as objects pairs (X, v) with $v \in G$, $v(Y) \leq f(X)$ and there is a morphism from (X, v) to (X', v') if and only if $v^{-1}(X) = v'^{-1}(X')$, in which case the morphism is unique. We define a functor g from $Y\backslash f$ to $f^{-1}(Y)$ on objects by $g((X, v)) = v^{-1}(X)$. It is easy to check that this extends in a unique way to morphisms. It is a right adjoint to f because $Y \leq g((X, v))$ if and only if there is a morphism (necessarily unique) from $f(Y)$ to (X, v), both conditions being equivalent to $v(Y) \leq X$. This shows that the conditions of Quillen's result are satisfied, so f is a homotopy equivalence. This proves lemma 3.2.

Lemma 3.3. *The functor f is a homotopy equivalence from Q_2 to Q_1.*

Proof. By lemma 3.2 it is sufficient to check that for each $Y \in P_1$, the sub poset $f^{-1}(Y)$ of P_2 is contractible. The poset $f^{-1}(Y)$ is the poset of a cell decomposition of a convex cone in a real vector space. As any convex set is contractible, the poset $f^{-1}(Y)$ is also contractible. This proves lemma 3.3.

Lemma 3.4. *Suppose (W, S) is a spherical Coxeter system acting on the vector space V_S with Weyl chamber C. Suppose K is an isometry from J into S. Let $V_S^{*W_K}$ be the subspace of V^* fixed by W_K, where W_K is the reflection group whose simple roots are the points $K(J)$. The walls of $C \cap V_S^{*W_K}$ correspond to the points in S not in the image of K;*

let s be one of these points and let $s^\perp \cap V_S^{*W_K}$ be the wall in $V_S^{*W_K}$ corresponding to s. Choose $w \in W$ so that $w(C)$ is the (unique) Weyl chamber of W such that $w(C) \cap V_S^{*W_K}$ is the cell in $V_S^{*W_K}$ on the other side of $s^\perp \cap V_S^{*W_K}$ to $C \cap V_S^{*W_K}$ and such that C and $w(C)$ are both in the same Weyl chamber of W_K. Then

$$w = \sigma_{K(J)\cup s}\sigma_{K(J)}.$$

Proof. We can reduce to the case when $S = K(J)\cup s$, so that $V_S^{*W_K}$ is one dimensional and $s^\perp \cap V_S^{*W_K}$ is just the point 0. Then $\sigma_S(\sigma_{K(J)}(C)) = -\sigma_{K(J)}(C)$ which contains $-C\cap V_S^{*W_K}$, so $\sigma_S(\sigma_{K(J)}(C))\cap V_S^{*W_K}$ is a cell on the other side of $s^\perp\cap V_S^{*W_K}$ to $C\cap V_S^{*W_K}$. Moreover $\sigma_{K(J)}(C)$ is in the opposite Weyl chamber of $K(J)$ to C, and $\sigma_S(\sigma_{K(J)}(C))$ is in the opposite Weyl chamber to $\sigma_{K(J)}(C)$, so $\sigma_S(\sigma_{K(J)}(C))$ is in the same Weyl chamber of $K(J)$ as C. This shows that the element w of the lemma is $\sigma_S\sigma_{K(J)}$. This proves lemma 3.4.

A result similar to the following lemma (using subsets of S rather than isometries $K : J \mapsto S$) is given in [H, lemma 5] when S is finite and in [D] for arbitrary S.

Lemma 3.5. *Suppose (W, S) is a Coxeter system, J is a spherical Coxeter diagram, and K and K' are two isometries from J into S. Then K and K' are conjugate under W if and only if they are associate.*

Proof. First suppose that K and K' are adjacent. Then $\sigma_{K\cup K'}(\sigma_K(K)) = K'$, so K and K' are conjugate under W. Next suppose K and K' are associate. Then by definition we can find a sequence $K = K_1, K_2, \ldots, K_n = K'$ such that K_i and K_{i+1} are adjacent for all i. Hence $K = K_1$ and $K' = K_n$ are also conjugate under W. So associate isometries from J into S are conjugate under W.

Conversely, suppose that K and K' are conjugate by an element $w \in W$. Consider the subspace $V_S^{*W_K}$ of V_S^*, and the codimension 0 cells in it of the form $V_S^{*W_K} \cap C$ for some Weyl chamber C of W. For any two such cells, for example $D = V_S^{*W_K} \cap C$ and $D' = V_S^{*W_K} \cap w(C)$, we can find a sequence $D = D_1, D_2, \ldots, D_n = D'$ such that D_i and D_{i+1} are adjacent by a face of codimension 1 in $V_S^{*W_K}$. For each i let C_i be the (unique) Weyl chamber whose intersection with $V_S^{*W_K}$ is D_i and that is contained in the Weyl chamber of W_K. We identify each C_i with C_Π using w_i. The set $K(J)$ is a subset of the simple roots of C_i, so $K_i = w_i^{-1}(K)$ maps J to the simple roots of C_Π. The isometries K_i and K_{i+1} are adjacent for all i, because the element $w_{i+1}w_i^{-1}$ mapping C_i to C_{i+1} is equal to $\sigma_{K(J)\cup s_i}\sigma_{K(J)}$, where s_i is the simple root of C_i orthogonal to $D_i \cap D_{i+1}$ but not to D_i. Therefore $K = K_1$ and $K' = K_n$ are adjacent. This proves lemma 3.5.

Suppose that P is a W-poset for a group W with the property that if $p \le w(p)$ for $w \in W, p \in P$ then $p = w(p)$. We define the **quotient** $W\backslash P$ of P by W to be the poset whose elements are the orbits Wp of W acting on P, where we put $Wp \le Wq$ if $w(p) \le q$ for some $w \in W$. This should not be confused with the homotopy quotient of P by W.

Lemma 3.6. *The $\Gamma_J \times \Gamma_\Pi$ posets P_3^+ and $W_\Pi\backslash P_2^+$ are isomorphic.*

Proof. We will construct an isomorphism f of posets from $W_\Pi\backslash P_2^+$ to P_3^+. Suppose (D, K) is an element of P_2^+ representing an element of $W_\Pi\backslash P_2^+$. We can find an element

w of W_Π such that $w(D) \subseteq C_\Pi$ and $w(K(J)) \subseteq \Pi$. We define $f((D,K))$ to be $(S, \overline{w(K)})$, where S is the set of simple roots of C_Π orthogonal to $w(D)$. We check that this is well defined even though w is not unique. To prove this we can assume that $D \subseteq C_\Pi$ and $K(J) \subseteq \Pi$. Then the different possibilities for w are elements of the group generated by the reflections fixing D and the Weyl chamber of $K(J)$. But these elements take K to an associated isometry $K : J \mapsto S$, so the equivalence class $\overline{K}$ is well defined by lemma 3.5. These elements also take S to S, so S is well defined. This proves that $(S, \overline{K})$ is uniquely defined.

The isomorphism f of posets from $W_\Pi \backslash P_2^+$ to P_3^+ obviously preserves the $\Gamma_J \times \Gamma_\Pi$ action on both posets. This proves lemma 3.6.

Lemma 3.7. *Suppose (W,S) is a spherical Coxeter system, K is an isometry $K : J \mapsto S$, and s is a point of S not in $K(J)$. Then there is an element of W mapping $V_S^{*W_K}$ to itself and acting on $V_S^{*W_K}$ as reflection in $s^\perp \cap V_S^{*W_K}$ if and only if $-\sigma_{K(J)\cup s}$ maps $K(J)$ to itself. If such an element of W exists, then there is a unique such element w mapping K to itself, given by $w = \sigma_{K(J)\cup s}\sigma_{K(J)}$.*

Proof. If an element of w maps $V_S^{*W_K}$ to itself then there is a unique element of W with the same action on $V_S^{*W_K}$ and mapping $K(J)$ to itself because W_K acts simply transitively on its Weyl chambers, so we may assume that w maps $K(J)$ to itself. If in addition w acts on $V_S^{*W_K}$ as reflection in $s^\perp \cap V_S^{*W_K}$ then by lemma 3.4 w must be $\sigma_{K(J)\cup s}\sigma_{K(J)}$.

Conversely if $-\sigma_{K(J)\cup s}$ maps $K(J)$ to itself then $\sigma_{K(J)\cup s}\sigma_{K(J)}$ maps $K(J)$ to itself and acts on $V_S^{*W_K}$ as reflection in $s^\perp \cap V_S^{*W_K}$. This proves lemma 3.7.

Lemma 3.8. *Suppose K is an isometry from J into a spherical subdiagram S of Π. Then $K : J \mapsto S$ is R-reflective if and only if the Π-cell $S^\perp \cap C_\Pi$ is contained in a reflection hyperplane of W_{Ω_K} of the form $s^\perp \cap V_\Pi^{*W_K}$ for $s \in \Pi$.*

Proof. First suppose that $K : J \mapsto S$ is R-reflective. Then there is a point $s \in S$ not in $K(J)$ such that $w = \sigma_{K(J)\cup s}\sigma_{K(J)}$ acts on $K(J)$ as an element of R. So w is a reflection of W_Ω corresponding to the hyperplane $s^\perp \cap V_\Pi^{*W_K}$, and this hyperplane contains $S^\perp \cap C_\Pi$.

Conversely, suppose that $S^\perp \cap C_\Pi$ is contained in a reflection hyperplane of $w \in W_{\Omega_K}$ of the form $s^\perp \cap V_\Pi^{*W_K}$ for $s \in \Pi$. Then we must have $s \in S$ because $S^\perp \cap C_\Pi \subseteq s^\perp$. The element w must be equal to $\sigma_{K(J)\cup s}\sigma_{K(J)}$, and this element acts on K as an element of R because $w \in W_\Omega$. Therefore $K : J \mapsto S$ is R-reflective. This proves lemma 3.8.

Lemma 3.9. *Suppose K is an isometry from J into a spherical subdiagram S of Π. Then the equivalence class $\overline{K}$ is R-reflective if and only if the Π-cell $S^\perp \cap C_\Pi$ of $V_\Pi^{*W_K}$ corresponding to S is contained in a reflection hyperplane of W_{Ω_K}.*

Proof. Suppose the cell of $V_\Pi^{*W_K}$ corresponding to S is contained in a reflection hyperplane of W_{Ω_K}. Choose a Weyl chamber C_Π for W_Π such that this reflection hyperplane is a wall of $C_\Pi \cap V_\Pi^{*W_K}$. Then by lemma 3.8 the corresponding isometry $K' : J \mapsto S$ is R-reflective, and by lemma 3.5 is associate to K. So $\overline{K}$ is R-reflective.

Conversely suppose that $\overline{K}$ is R-reflective. Then $w(K)$ is R-reflective and has image in S for some $w \in W_\Omega$. By lemma 3.8 this implies that $w(S^\perp \cap C_\Pi)$ is contained in a reflection hyperplane of W_Ω, so the same is true of $S^\perp \cap C_\Pi$. This proves lemma 3.9.

Lemma 3.10. *Suppose P is a G-poset for some group G. Suppose that W is a normal subgroup of G such that if $p \leq q$ and $w(p) \leq q$ for some $w \in W$, $p, q \in P$, then $w = 1$. Then the homotopy quotient Q_2 of P by G is equivalent to the homotopy quotient Q_3 of $W\backslash P$ by G/W.*

Proof. Recall from [M, theorem 1, page 91] that if f is any functor from a category Q_2 to a category Q_3 then f is an equivalence if the following two conditions are satisfied:

1 Any object of Q_3 is isomorphic to some object in the image of f.
2 For any two objects p, q of Q_2, f induces an isomorphism from $Mor(p, q)$ to $Mor(f(p), f(q))$.

We will apply this to show that our categories Q_2 and Q_3 are equivalent. We define f on objects by $f(p) = Wp$, and define f on morphisms using the obvious homomorphism from G to G/W. Condition 1 above is satisfied because every element of P_3 is the image of an element of P_2, so every object of Q_3 is the image of an object of Q_2. Suppose p and q are objects of P_2. The set of morphisms of Q_2 from p to q can be identified with the set of elements g of G such that $g(p) \leq q$, and $Mor_{Q_3}(f(p), f(q))$ can also be identified with the set of elements g of G such that $g(p) \leq q$, so condition 2 above is satisfied. This shows that f is an equivalence and proves lemma 3.10.

Lemma 3.11. *If $p \leq q$ and $w(p) \leq q$ for some $w \in W_\Pi$, $p, q \in P_2$, then $w = 1$.*

Proof. Suppose that $p = (D, K)$. Then $q = (D_1, K)$ for some D_1 containing D as $p \leq q$. But then $w(D) \subseteq D_1$, so $w(D) = D$ as no two distinct subsets D, $w(D)$ of D_1 are conjugate under W_Π, as D_1 is contained in a fundamental domain of W_Π. Hence we can assume that w fixes D as well as K.

The subgroup of W_Π fixing D is a finite reflection group W_D generated by the reflections of W_Π fixing D because $D \subseteq C_\Pi$. The subgroup of W_Π fixing K is generated by the reflections fixing all elements of $K(J)$. Any such reflection is in W_{Ω_K} because R contains 1. However, the condition that $p = (D, K) \in P_2$ implies that there are no reflections of W_{Ω_K} fixing D. Hence the subgroup of W_D fixing K is trivial. So any element $w \in W_\Pi$ such that $w(p) \leq q$ is trivial. This proves lemma 3.11.

Lemma 3.12. *The categories Q_2 and Q_3 are equivalent.*

Proof. Lemmas 3.10 and 3.11 show that there is an equivalence of categories from Q_2 to the homotopy quotient of $W_\Pi\backslash P_2$ by $\Gamma_J \times \Gamma_\Pi$. Lemma 3.6 shows that the $\Gamma_J \times \Gamma_\Pi$ posets P_3^+ and $W_\Pi\backslash P_2^+$ are isomorphic. Lemma 3.9 shows that the subset P_2 of P_2^+ corresponds under this isomorphism to the subset $W_\Pi\backslash P_2$ of $W_\Pi\backslash P_2^+$, so the $\Gamma_J \times \Gamma_\Pi$ posets P_3 and $W_\Pi\backslash P_2$ are isomorphic. Therefore the category Q_2 is equivalent to the homotopy quotient of P_3 by $\Gamma_J \times \Gamma_\Pi$, which is just Q_3. This proves lemma 3.12.

Lemma 3.13. *The natural injection from Q_4 to Q_3 is an equivalence of categories.*

Proof. This follows from the fact that Q_4 is a skeleton of Q_3, so the natural injection is an equivalence of categories. This proves lemma 3.13.

We can now prove theorem 2.7. By lemmas 3.3, 3.12, and 3.13, the categories Q_4 and Q_1 are homotopy equivalent. So by lemma 3.1, the component of Q_4 containing $(J, \overline{id_J})$ is a classifying category for Γ_Ω. This proves theorem 2.7.

4. The structure of Γ_Ω.

Theorem 4.1. *The kernel of the natural map from Γ_Ω to $\Gamma_J \times \Gamma_\Pi$ has finite cohomological dimension.*

Proof. The classifying category of the kernel is just a component of the category of the poset P_3. A case by case check on possible Coxeter diagrams shows that the lengths of chains in P_3 are bounded (by $rank(J)+1$ for example), so the corresponding simplicial complex has finite dimension at most $rank(J)$. Therefore the kernel has cohomological dimension at most $rank(J)$. This proves theorem 4.1.

Corollary 4.2. *If Γ_Π has finite virtual cohomological dimension, then so does Γ_Ω.*

Proof. This follows immediately from theorem 4.1 and the fact that Γ_J is finite and standard properties of the virtual cohomological dimension.

Howlett showed that if W_Π is finite then the group Γ_Ω is a subgroup of $Aut(J) \times Aut(\Pi)$. We can deduce this from theorem 4.1 as follows. If W_Π is finite then so is the kernel of the map from Γ_Ω to $Aut(J) \times Aut(\Pi)$. On the other hand this kernel has finite cohomological dimension by theorem 4.1. But any finite group of finite cohomological dimension must be trivial, so the natural map from Γ_Ω to $Aut(J) \times Aut(\Pi)$ is injective. If W_Π is infinite then this kernel is usually infinite, as can be seen from most of the examples below. The fact that this kernel no longer vanishes is the main reason why normalizers of parabolic subgroups of Coxeter groups are more complicated to describe when the Coxeter group is infinite.

5. Examples.

Example 5.1. Suppose that J is A_1 and the group Γ_Π is trivial. In this case Brink [Br] gave an elegant description of the group Γ_Ω as follows. Form the graph obtained from the Coxeter graph Π by keeping only the edges of odd order. Then for any point $J = A_1$ of this new graph, the centralizer of the corresponding reflection (which is the group Γ_Ω corresponding to J) is the fundamental group of this graph with basepoint the chosen point, and in particular Γ_Ω is a free group.

We now check that this is equivalent to the description given by theorem 2.7. The only subdiagrams S with a non R-reflective class $\overline{K}$ are the points of the Coxeter graph or the edges of odd order together with their endpoints. So the category Q_4 has an object for each point or odd order edge of the Coxeter graph. The only non-identity morphisms correspond to inclusions of points in edges. The classifying space of this category is just the first barycentric subdivision of Brink's graph. In particular the fundamental group of this category with some object as basepoint is canonically isomorphic to the fundamental group of Brink's graph with the corresponding point as basepoint. This verifies that the description of Γ_Ω in [Br] is equivalent to the description in theorem 2.7.

The following lemma can often be used to find the fundamental group of a category with at most 2 objects.

Lemma 5.2. *Suppose A and B are subgroups of a group. Let Q be the category with 2 objects p and q such that $Mor(p,p) = A$, $Mor(p,q) = BA$, $Mor(q,q) = B$, $Mor(q,p) = \emptyset$, with composition defined in the obvious way. Then $\pi_1(Q) = A *_{A\cap B} B$.*

Proof. This can be proved by writing down a set of generators and relations for the fundamental group, and checking that they are equivalent to a set of generators and relations for $A *_{A\cap B} B$. We will leave the details to the reader.

For most of the examples below we will take $W_\Pi.\Gamma_\Pi$ to be the group of automorphisms of the even 26 dimensional Lorentzian lattice $II_{1,25}$ not exchanging the two cones of norm 0 vectors. According to Conway [C], the Coxeter group W_Π has a simple reflection $r_\lambda \in \Pi$ for each vector λ of the Leech lattice $\Lambda = \Pi$, and the order of $r_\lambda r_\mu$ is 1, 2, 3, or ∞ according to whether $(\lambda - \mu)^2$ is 0, 4, 6, or greater than 6. The group Γ_Π is the automorphism group $\Lambda.Aut(\Lambda) = \Lambda.(Z/2Z).Co_1$ of the affine Leech lattice, where Λ is the subgroup of translations and Co_1 is Conway's largest sporadic simple group.

If J is a spherical subdiagram of Λ then there is a homomorphism from $N_{W_\Pi.\Gamma_\Pi}(W_J)/W_J$ to the automorphism group of the lattice $J^\perp$. This has finite kernel and co-finite image, so theorem 2.7 can usually be used to describe the automorphism group of the lattice $J^\perp$. Most of the remaining examples in this section use this idea.

Example 5.3. Suppose L is the even Lorentzian lattice of dimension 20 and determinant 3. Let $W^{(2)}(L)$ be the subgroup of $Aut(L)$ generated by reflections of norm -2 vectors of L. Vinberg showed in [V] that $Aut(L)^+/W^{(2)}(L)$ was the automorphism group of a certain K3 surface modulo a cyclic subgroup, and also showed that this group was an extension of a group of order 72 by a free product of 12 groups of order 2. We will show how to recover Vinberg's description of $Aut(L)^+/W^{(2)}(L)$ from theorem 2.7.

We take J to be an $E_6 \subset \Lambda$ so that $L = J^\perp$, and take $R = 1 \subset \Gamma_J = Aut(E_6) = Z/2Z$. The category Q_4 contains exactly two objects, corresponding to an E_6 and an E_7 in Λ. For each subdiagram X of Λ we write $G(X)$ for the automorphisms of $\Pi = \Lambda$ mapping X into itself. The morphisms from the E_6 object to itself form a group $G(E_6)$ of order 72. The morphisms from the E_7 object to itself form a group $Z/2Z \times G(E_7)$ of order $2 \times 6 = 12$ (where the $Z/2Z$ comes from the group Γ_J). The morphisms from E_6 to E_7 can be identified with $Z/2Z \times G(E_6)$. By lemma 5.2 the group Γ_Ω is isomorphic to $G(E_6) *_{G(E_7)} (G(E_7) \times Z/2Z)$. If A, B, and C are any groups with $B \subseteq A$ then $A *_B (B \times C)$ is a semidirect product of a normal subgroup isomorphic to the free product of $|A|/|B|$ copies of C and with the quotient by this normal subgroup isomorphic to A. Hence Γ_Ω has a normal subgroup isomorphic to the free product of 12 copies of $Z/2Z$, and the quotient is the group $G(E_6)$ of order 72. This is equivalent to the description of this group given by Vinberg in [V]. In fact we can describe the various parts of Vinberg's description in terms of the Leech lattice as follows: the group of order 2 is the group of automorphisms of E_6, the group of order 72 is the subgroup of $Aut(\Lambda)$ mapping an E_6 into itself, and the number 12 is the number of E_7's of Λ containing an E_6.

The group $Aut(L)^+$ also contains reflections in norm -6 vectors. The quotient by the full reflection group is finite of order 72, isomorphic to $G(E_6)$. In this case the category Q_4 contains just one point. Note that the reflections of norm -6 vectors induce the nontrivial automorphism of E_6. The 12 elements of order 2 in the paragraph above are in fact reflections of norm -6 vectors. See [V] or [B] for more details of this case.

If L is the even Lorentzian lattice of determinant 4 and dimension 20, which is again the Picard lattice of a $K3$ surface, then Vinberg gave a similar description of the automorphism group as an extension $((Z/2Z) * (Z/2Z) * (Z/2Z) * (Z/2Z) * (Z/2Z)).S_5$ of the symmetric

group S_5 by the free product of 5 group of order 2. This group can also be calculated using theorem 2.7. The corresponding category Q_4 has 2 objects, corresponding to a D_6 or D_7 in Λ, the group S_5 is the subgroup of $Aut(\Lambda)$ mapping the D_6 to itself, the group $Z/2Z$ is the group of automorphisms of the D_6 diagram, and the number 5 of copies of $Z/2Z$ is the number of D_7's containing a D_6.

Example 5.4 As a more complicated example we will describe the automorphism group of the Picard lattice L of the "next most algebraic K3 surface", in other words L is the 20 dimensional even Lorentzian lattice of determinant 7. We take $\Pi = \Lambda$, $\Gamma_\Pi = \Lambda.Aut(\Lambda)$, $J = A_6$, $R = 1$, $\Gamma_J = Aut(A_6) = Z/2Z$. Then Γ_Ω is the subgroup of elements of $Aut(L)$ fixing a Weyl chamber of the reflection group generated by the reflections of norm -2 vectors. By theorem 2.7 the group Γ_Ω is the fundamental group of the category Q_4. The category Q_4 has exactly 5 objects, corresponding to the 5 orbits of Coxeter diagrams A_6, A_7, D_7, E_7, and D_8 with a non R-reflective isometry from A_6 into them. (The group $\Lambda.Aut(\Lambda)$ acts transitively on the embeddings of any of these Coxeter diagrams into Λ.) Note that for D_8 and E_7 there is only one equivalence class of isometries from $A_6 = J$ into it, while for A_6, A_7, and D_7 there are two classes, which are exchanged by $Aut(J) = Z/2Z$. The category Q_4 looks like this.

$$\begin{array}{ccccccc} & & & & A_7(48) & & \\ & & (672) \nearrow & & & \searrow (192) & \\ E_7(12) & \longleftarrow & A_6(336) & \longrightarrow & & D_8(16) & \\ & (672) & (672) \searrow & & (2016) & \nearrow (48) & \\ & & & & D_7(24) & & \end{array}$$

Here the numbers are the numbers of morphisms between pairs of objects in Q_4.

Example 5.5. Kondo in [K] studied the automorphism group of a generic Jacobian Kummer surface by embedding its Picard lattice L as the orthogonal complement of a certain $J = A_3A_1^6$ in $II_{1,25}$, and used this to describe a generating set for the automorphism group. (Note that the Leech lattice contains more than 1 orbit of subdiagrams of the form $A_3A_1^6$; the one used by Kondo has largest possible stabilizer in $Aut(\Lambda)$.) By results of Nikulin [N] the automorphism group of the K3 surface is the subgroup of $Aut(L)^+/W^{(2)}(L)$ of elements acting on L'/L as ± 1. This is just the group Γ_Ω of theorem 2.7 where we take $\Pi = \Lambda$, $\Gamma_\Pi = \Lambda.Aut(\Lambda)$, J to be Kondo's $A_3A_1^6$, $R = 1$, and $\Gamma_J =$ the subgroup of order 2 of $Aut(J)$ generated by the nontrivial automorphism of the A_3. The category Q_4 is not connected and the component containing J seems quite complicated. Some partial calculations I have done suggest that the automorphism group of the generic Jacobian Kummer surface might be

$$(W.(Z/2Z)^5) * (Z/2Z) * (Z/2Z) * (Z/2Z) * (Z/2Z) * (Z/2Z) * (Z/2Z)$$

where W is a Coxeter group of rank 32+60 generated by the 16 projections, 16 correlations, sixty Cremona transformations and the $(Z/2Z)^5$ is generated by sixteen translations and a switch [K], but I have not proved this rigorously.

It is much easier to work out the group Γ_Ω for R and Γ_J replaced by $Aut(J) = Z/2Z \times S_6$. In this case the group Γ_Ω has a normal subgroup of index $|S_6|$ isomorphic

to the quotient of the automorphism group of a generic Jacobian Kummer surface by the Coxeter group generated by reflections in norm -4 vectors. Theorem 2.7 describes the Γ_Ω as the fundamental group of a component of the finite category Q_4. This component has just two elements, corresponding to the Coxeter diagrams $A_3A_1^6$ and $A_5A_1^5$. Using lemma 5.2 and the results in [K] we see that

$$\Gamma_\Omega = ((Z/2Z)^5.S_6) *_{S_5} (S_5 \times Z/2Z)$$

where $(Z/2Z)^5.S_6$ is the subgroup of Γ_Π fixing $A_3A_1^6$ ([K, lemma 4.5]), and $S_5 \times Z/2Z$ is the subgroup of Γ_Π fixing $A_5A_1^5$. In particular Γ_Ω has a normal subgroup which is the free product of 192 groups of order 2, and the quotient by this normal subgroup is $(Z/2Z)^5.S_6$. If we change R to 1 but keep $\Gamma_J = Aut(J)$ then Γ_Ω becomes the group $Aut(L)^+/W^{(2)}(L)$ which appears to be

$$(W.(Z/2Z)^5.S_6) *_{S_5} (S_5 \times (Z/2Z))$$

though I have not proved this rigorously.

For more examples of automorphism groups of Kummer surfaces, corresponding to the cases $J = D_4^2$, D_4A_3, D_4A_2, or A_3^2, see [K-K].

Example 5.6 Suppose that J in theorem 2.7 contains no components of types A_n ($n \geq 1$) or D_5, and assume that $R = \Gamma_J = Aut(J)$. Then the map from Γ_Ω to $\Gamma_J \times \Gamma_\Pi$ is injective. This follows because a case by case check over all irreducible spherical Coxeter diagrams shows that any isometry from J into a strictly larger spherical Coxeter diagram is R-reflective.

Example 5.7 We show how to explain Vinberg's result [V, V-K] that the reflection group of $I_{1,n}$ has finite index if and only if $n \leq 19$. Following Conway and Sloane [C-S] we write the even sublattice L of $I_{1,n}$ as $D_{25-n}^\perp$ in $II_{1,25}$ for $n \leq 23$, where $D_3 = A_3$ and $D_2 = A_1^2$. In theorem 2.7 we take $\Pi = \Lambda$, $\Gamma_\Pi = \Lambda.Aut(\Lambda)$, $J = D_{25-n}$, $R = \Gamma_J$ a subgroup of order 2 of $Aut(J)$ (which is equal to $Aut(J)$ for $n \neq 21$). Then the quotient of $Aut(L)^+$ by its reflection subgroup is the group Γ_Ω. For $n \leq 19$ the group Γ_Ω is finite by example 5.6. For $n = 20$ this argument breaks down because J is the "exceptional" case D_5 of example 5.6. For $20 \leq n \leq 23$ we can still describe the group Γ_Ω explicitly using theorem 2.7; see [B, theorem 6.6] for details. When $n = 21$ this gives a natural example with $\Gamma_J \neq Aut(J)$.

If we take J to be D_4 and take $R = \Gamma_J$ to be the symmetric group $S_3 = Aut(D_4)$ instead of a group of order 2 then $W_\Omega.\Gamma_\Omega$ is the automorphism group of the even sublattice of $I_{1,21}$, and Γ_Ω is a finite group. See [B, p. 149] for details.

Example 5.8. The groups Γ_Ω have many of the properties of arithmetic groups; for example, they often have finite classifying categories. (It follows easily from [S] that arithmetic groups have this property.) It is natural to ask when they are arithmetic. There seems to be no obvious general way of deciding this. The following argument can often be used to show that Γ_Ω is not arithmetic. First of all a theorem due to Margulis [Ma, page 3] implies that if a group is an arithmetic subgroup of a group of rank at least 2, then all normal subgroups are either in the center or of finite index. Secondly, a theorem of Borel and Serre [B-S, 11.4.4] says that if a group Γ is arithmetic in a Lie group G then $d = r + vcd(\Gamma)$ where d is the dimension of the symmetric space of G and r is the rank

of G and $vcd(\Gamma)$ is the virtual cohomological dimension of Γ. Let us use these results to prove that the group Γ_Ω of example 5.3 (the automorphism group of a K3 surface) is not arithmetic in any Lie group G. The group G must have rank 1 by the theorem of Margulis, as Γ_Ω has non abelian free subgroups of finite index and so cannot be an arithmetic subgroup of a group of rank at least 2. Its virtual cohomological dimension is one, so by the theorem of Borel and Serre [B-S, 11.4.4] the symmetric space of G must have dimension $1+1=2$. But any finite subgroup of Γ_Ω must fix a point of this symmetric space, and therefore acts faithfully on the 2 dimensional tangent space of this point. But Γ_Ω has finite subgroups that are too large to have 2 dimensional faithful representations. Hence Γ_Ω is not arithmetic. Note that Γ_Ω has subgroups of finite index that are free and therefore arithmetic.

References.

[B] R. E. Borcherds, Automorphism groups of Lorentzian lattices, J.Alg. Vol 111 (1987) No. 1, pp. 133-153.

[B-S] A. Borel, J.-P. Serre, Corners and arithmetic groups. Comment. Math. Helv. 48 (1973), 436–491.

[B-H] B. Brink, R. B. Howlett, Normalizers of Parabolic Subgroups in Coxeter Groups, 1998 preprint, available from http://www.maths.usyd.edu.au:8000/u/~bobh/

[Bo] N. Bourbaki, Groupes et algèbres de Lie. Chapitres IV-VI. Actualités Scientifiques et Industrielles, No. 1337 Hermann, Paris 1968.

[Br] B. Brink, On centralizers of reflections in Coxeter groups. Bull. London Math. Soc. 28 (1996), no. 5, 465–470.

[C] J. H. Conway, The automorphism group of the 26-dimensional even unimodular Lorentzian lattice. J. Algebra 80 (1983), no. 1, 159–163. Reprinted as chapter 27 of "Sphere packings, lattices and groups." Grundlehren der Mathematischen Wissenschaften 290. Springer-Verlag, New York, 1993. ISBN: 0-387-97912-3

[C-S] J. H. Conway, N. J. A. Sloane, Leech roots and Vinberg groups. Proc. Roy. Soc. London Ser. A 384 (1982), no. 1787, 233–258. Reprinted as chapter 28 of "Sphere packings, lattices and groups." Grundlehren der Mathematischen Wissenschaften 290. Springer-Verlag, New York, 1993. ISBN: 0-387-97912-3

[D] V. V. Deodhar, On the root system of a Coxeter group, Comm. Algebra 10 (1982) 611–630.

[Hi] H. Hiller, Geometry of Coxeter groups. Research Notes in Mathematics, 54. Pitman (Advanced Publishing Program), Boston, Mass.-London, 1982. ISBN: 0-273-08517-4

[H] R. B. Howlett, Normalizers of parabolic subgroups of reflection groups. J. London Math. Soc. (2) 21 (1980), no. 1, 62–80.

[Ke] J. H. Keum, Automorphisms of Jacobian Kummer surfaces. Compositio Math. 107 (1997), no. 3, 269–288.

[K-K] J. H. Keum, S. Kondo, The automorphism groups of Kummer surfaces associated with the product of two elliptic curves, 1998 preprint.

[K] S. Kondo, The automorphism group of a generic Jacobian Kummer surface. J. Algebraic Geometry 7 (1998) 589–609.

[K98] S. Kondo, Niemeier lattices, Mathieu groups and finite groups of symplectic automorphisms of K3 surfaces. Duke Math. Journal Vol 92 (1998) no. 3, 593-598.

[M] S. MacLane, Categories for the working mathematician. Graduate Texts in Mathematics, Vol. 5. Springer-Verlag, New York-Berlin, 1971.

[Ma] G. A. Margulis, "Discrete subgroups of semisimple Lie groups". Ergebnisse der Mathematik und ihrer Grenzgebiete (3), 17. Springer-Verlag, Berlin, 1991. ISBN: 3-540-12179-X.

[Mu] S. Mukai, Finite groups of automorphisms of K3 surfaces and the Mathieu group. Invent. Math. 94 (1988), no. 1, 183–221.

[N] V. V. Nikulin, An analogue of the Torelli theorem for Kummer surfaces of Jacobians. (Russian) Izv. Akad. Nauk SSSR Ser. Mat. 38 (1974), 22–41. Integral symmetric bilinear forms and its applications, Math. USSR Izv., 14 (1980), 103-167.

[P-S] I. R. Shafarevich, I. I. Piatetski-Shapiro, Torelli's theorem for algebraic surfaces of type K3. (Russian) Izv. Akad. Nauk SSSR Ser. Mat. 35 1971 530–572. English translation in pages 516-557 of Collected mathematical papers by Igor R. Shafarevich. Springer-Verlag, Berlin-New York, 1989. ISBN: 3-540-13618-5

[Q] D. Quillen, Higher algebraic K-theory. I. Algebraic K-theory, I: Higher K-theories (Proc. Conf., Battelle Memorial Inst., Seattle, Wash., 1972), pp. 85–147. Lecture Notes in Math., Vol. 341, Springer, Berlin 1973.

[S] J.-P. Serre, Cohomologie des groupes discrets. Prospects in mathematics (Proc. Symp., Princeton Univ., Princeton, N.J., 1970), pp. 77–169. Ann. of Math. Studies, No. 70, Princeton Univ. Press, Princeton, N.J., 1971.

[V] È. B. Vinberg, The two most algebraic $K3$ surfaces. Math. Ann. 265 (1983), no. 1, 1–21.

[V-K] È. B. Vinberg, I. M. Kaplinskaja, The groups $O_{18,1}(Z)$ and $O_{19,1}(Z)$. (Russian) Dokl. Akad. Nauk SSSR 238 (1978), no. 6, 1273–1275.

arXiv:math/9809110v1 [math.QA] 19 Sep 1998

What is moonshine?

Richard E. Borcherds, *
D.P.M.M.S., 16 Mill Lane, Cambridge, CB2 1SB, England.
e-mail: reb@dpmms.cam.ac.uk
home page: www.dpmms.cam.ac.uk/~reb

This is an informal write up of my talk at the I.C.M. in Berlin. It gives some background to Goddard's talk [Go] about the moonshine conjectures. For other survey talks about similar topics see [B94], [B98], [LZ], [J], [Ge], [Y].

The classification of finite simple groups shows that every finite simple group either fits into one of about 20 infinite families, or is one of 26 exceptions, called sporadic simple groups. The monster simple group is the largest of the sporadic finite simple groups, and was discovered by Fischer and Griess [G]. Its order is

$$\begin{aligned}&8080,17424,79451,28758,86459,90496,17107,57005,75436,80000,00000\\ =&2^{46}.3^{20}.5^{9}.7^{6}.11^{2}.13^{3}.17.19.23.29.31.41.47.59.71\end{aligned}$$

(which is roughly the number of elementary particles in the earth). The smallest irreducible representations have dimensions $1, 196883, 21296876, \ldots$. The elliptic modular function $j(\tau)$ has the power series expansion

$$j(\tau) = q^{-1} + 744 + 196884q + 21493760q^2 + \ldots$$

where $q = e^{2\pi i\tau}$, and is in some sense the simplest nonconstant function satisfying the functional equations $j(\tau) = j(\tau + 1) = j(-1/\tau)$. John McKay noticed some rather weird relations between coefficients of the elliptic modular function and the representations of the monster as follows:

$$\begin{aligned}1 &= 1\\ 196884 &= 196883 + 1\\ 21493760 &= 21296876 + 196883 + 1\end{aligned}$$

where the numbers on the left are coefficients of $j(\tau)$ and the numbers on the right are dimensions of irreducible representations of the monster. At the time he discovered these relations, several people thought it so unlikely that there could be a relation between the monster and the elliptic modular function that they politely told McKay that he was talking nonsense. The term "monstrous moonshine" (coined by Conway) refers to various extensions of McKay's observation, and in particular to relations between sporadic simple groups and modular functions.

For the benefit of readers who are not native English speakers, I had better point out that "moonshine" is not a poetic terms referring to light from the moon. It means foolish or crazy ideas. (Quatsch in German.) A typical example of its use is the following

* Supported by a Royal Society professorship.

quotation from E. Rutherford (the discoverer of the nucleus of the atom): "The energy produced by the breaking down of the atom is a very poor kind of thing. Anyone who expects a source of power from the transformations of these atoms is talking moonshine." (Moonshine is also a name for corn whiskey, especially if it has been smuggled or distilled illegally.)

We recall the definition of the elliptic modular function $j(\tau)$. The group $SL_2(\mathbf{Z})$ acts on the upper half plane H by

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}(\tau) = \frac{a\tau + b}{c\tau + d}.$$

A modular function (of level 1) is a function f on H such that $f((a\tau+b)/(c\tau+d)) = f(\tau)$ for all $\binom{ab}{cd} \in SL_2(\mathbf{Z})$. It is sufficient to assume that f is invariant under the generators $\tau \mapsto \tau + 1$ and $\tau \mapsto -1/\tau$ of $SL_2(\mathbf{Z})$. The elliptic modular function j is the simplest nonconstant example, in the sense that any other modular function can be written as a function of j. It can be defined as follows:

$$\begin{aligned} j(\tau) &= \frac{E_4(\tau)^3}{\Delta(\tau)} \\ &= q^{-1} + 744 + 196884q + 21493760q^2 + \cdots \\ E_4(\tau) &= 1 + 240 \sum_{n>0} \sigma_3(n) q^n \qquad\qquad (\sigma_3(n) = \sum_{d|n} d^3) \\ &= 1 + 240q + 2160q^2 + \cdots \\ \Delta(\tau) &= q \prod_{n>0} (1-q^n)^{24} \\ &= q - 24q + 252q^2 + \cdots \end{aligned}$$

A modular form of weight k is a holomorphic function $f(\tau) = \sum_{n\geq 0} c(n)q^n$ on the upper half plane satisfying the functional equation $f((a\tau + b)/(c\tau + d)) = (c\tau + d)^k f(\tau)$ for all $\binom{ab}{cd} \in SL_2(\mathbf{Z})$. The function $E_4(\tau)$ is an Eisenstein series and is a modular form of weight 4, while $\Delta(\tau)$ is a modular form of weight 12.

The function $j(\tau)$ is an isomorphism from the quotient $SL_2(\mathbf{Z})\backslash H$ to $\mathbf{C}$, and is uniquely defined by this up to multiplication by a constant or addition of a constant. In particular any other modular function is a function of j, so j is in some sense the simplest nonconstant modular function.

An amusing property of j (which so far seems to have no relation with moonshine) is that $j(\tau)$ is an algebraic integer whenever τ is an imaginary quadratic irrational number. A well known consequence of this is that

$$\exp(\pi\sqrt{163}) = 262537412640768743.99999999999925\ldots$$

is very nearly an integer. The explanation of this is that $j((1 + i\sqrt{163})/2)$ is exactly the integer $-262537412640768000 = -2^{18}3^35^323^329^3$, and

$$\begin{aligned} j((1 + i\sqrt{163})/2) &= q^{-1} + 744 + 196884q + \cdots \\ &= -e^{\pi\sqrt{163}} + 744 + (\text{something very small}). \end{aligned}$$

McKay and Thompson suggested that there should be a graded representation $V = \oplus_{n\in\mathbf{Z}} V_n$ of the monster, such that $\dim(V_n) = c(n-1)$, where $j(\tau) - 744 = \sum_n c(n)q^n = q^{-1} + 196884q + \cdots$. Obviously this is a vacuous statement if interpreted literally, as we could for example just take each V_n to be a trivial representation. To characterize V, Thompson suggested looking at the McKay-Thompson series

$$T_g(\tau) = \sum_n Tr(g|V_n)q^{n-1}$$

for each element g of the monster. For example, $T_1(\tau)$ should be the elliptic modular function. Conway and Norton [C-N] calculated the first few terms of each McKay-Thomson series by making a reasonable guess for the decomposition of the first few V_n's into irreducible representations of the monster. They discovered the astonishing fact that all the McKay-Thomson series appeared to be Hauptmoduls for certain genus 0 subgroups of $SL_2(\mathbf{Z})$. (A Hauptmodul for a subgroup Γ is an isomorphism from $\Gamma\backslash H$ to $\mathbf{C}$, normalized so that its Fourier series expansion starts off $q^{-1} + O(1)$.)

As an example of some Hauptmoduls of elements of the monster, we will look at the elements of order 2. There are 2 conjugacy classes of elements of order 2, usually called the elements of types $2A$ and $2B$. The corresponding McKay-Thompson series start off

$$T_{2B}(\tau) = q^{-1} + 276q - 2048q^2 + \cdots \qquad \text{Hauptmodul for } \Gamma_0(2)$$
$$T_{2A}(\tau) = q^{-1} + 4372q + 96256q^2 + \cdots \qquad \text{Hauptmodul for } \Gamma_0(2)+$$

The group $\Gamma_0(2)$ is $\{\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in SL_2(\mathbf{Z}) | c \text{ is even}\}$, and the group $\Gamma_0(2)+$ is the normalizer of $\Gamma_0(2)$ in $SL_2(\mathbf{R})$. Ogg had earlier commented on the fact that the full normalizer $\Gamma_0(p)+$ of $\Gamma_0(p)$ for p prime is a genus 0 group if and only if p is one of the primes 2, 3, 5, 7, 11, 13, 17, 19, 23, 29, 31, 41, 47, 59, or 71 dividing the order of the monster.

Conway and Norton's conjectures were soon proved by O. L. Atkin, P. Fong, and S. D. Smith. The point is that to prove something is a virtual character of a finite group it is only necessary to prove a finite number of congruences. In the case of the moonshine module V, proving the existence of an infinite dimensional representation of the monster whose McKay-Thompson series are give Hauptmoduls requires checking a finite number of congruences and positivity conditions for modular functions, which can be done by computer.

This does not give an explicit construction of V, or an explanation about why the conjectures are true. Frenkel, Lepowsky, and Meurman managed to find an explicit construction of a monster representation $V = \oplus V_n$, such that $\dim(V_n) = c(n-1)$, and this module had the advantage that it came with some extra algebraic structure preserved by the monster. However it was not obvious that V satisfied the Conway-Norton conjectures. So the main problem in moonshine was to show that the monster modules constructed by Frenkel, Lepowsky and Meurman on the one hand, and by Atkin, Fong, and Smith on the other hand, were in fact the same representation of the monster.

Peter Goddard [Go] has given a description of the proof of this in his talk in this volume, so I will only give a quick sketch of this. The main steps of the proof are as follows:

1. The module V constructed by Frenkel, Lepowsky, and Meurman has an algebraic structure making it into a "vertex algebra". A detailed proof of this is given in [F-L-M].
2. Use the vertex algebra structure on V and the Goddard-Thorn no-ghost theorem [G-T] from string theory to construct a Lie algebra acted on by the monster, called the monster Lie algebra.
3. The monster Lie algebra is a "generalized Kac-Moody algebra" ([K90]); use the (twisted) Weyl-Kac denominator formula to show that $T_g(\tau)$ is a "completely replicable function".
4. Y. Martin [M], C. Cummins, and T. Gannon [C-G] proved several theorems showing that completely replicable functions were modular functions of Hauptmoduls for genus 0 groups. By using these theorems it follows that T_g is a Hauptmodul for a genus 0 subgroup of $SL_2(\mathbf{Z})$, and hence V satisfies the moonshine conjectures. (The original proof used an earlier result by Koike [Ko] showing that the appropriate Hauptmoduls were completely replicable, together with a boring case by case check and the fact that a completely replicable function is characterized by its first few coefficients.)

We will now give a brief description of some of the terms above, starting with vertex algebras. The best reference for finding out more about vertex algebras is Kac's book [K]. In this paragraph we give a rather vague description. Suppose that V is a commutative ring acted on by a group G. We can form expressions like

$$u(x)v(y)w(z)$$

where $u, v, w \in V$ and $x, y, z \in G$, and the action of $x \in G$ on $u \in V$ is denoted rather confusingly by $u(x)$. (This is not a misprint for $x(u)$; the reason for this strange notation is to make the formulas compatible with those in quantum field theory, where u would be a quantum field and x a point of space-time.) For each fixed $u, v, \ldots \in V$, we can think of $u(x)v(y)\cdots$ as a function from G^n to V. We can rewrite the axioms for a commutative ring acted on by G in terms of these functions. We can now think of a vertex algebra roughly as follows: we are given lots of functions from G^n to V satisfying the axioms mentioned above, with the difference that these functions are allowed to have certain sorts of singularities. In other words a vertex algebra is a sort of commutative ring acted on by a group G, except that the multiplication is not defined everywhere but has singularities. In particular we cannot recover an underlying ring by defining the product of u and v to be $u(0)v(0)$, because the function $u(x)v(y)$ might happen to have a singularity at $u = v = 0$.

It is easy to write down examples of vertex algebras: any commutative ring acted on by a group G is an example. (Actually this is not quite correct: for technical reasons we should use a formal group G instead of a group G.) Conversely any vertex algebra "without singularities" can be constructed in this way. Unfortunately there are no easy examples of vertex algebras that are not really commutative rings. One reason for this is that nontrivial vertex algebras must be infinite dimensional; the point is that if a vertex algebra has a nontrivial singularity, then by differentiating it we can make the singularity worse and worse, so we must have an infinite dimensional space of singularities. This is only possible if the vertex algebra is infinite dimensional. However there are plenty of important infinite dimensional examples; see for example Kac's book for a construction of the most important examples, and [FLM] for a construction of the monster vertex algebra.

Next we give a brief description of generalized Kac-Moody algebras. The best way to think of these is as infinite dimensional Lie algebras which have most of the good properties of finite dimensional reductive Lie algebras. Consider a typical finite dimensional reductive Lie algebra G, (for example the Lie algebra $G = M_n(\mathbf{R})$ of $n \times n$ real matrices). This has the following properties:

1. G has an invariant symmetric bilinear form $(,)$ (for example $(a, b) = -Tr(a, b)$).
2. G has a (Cartan) involution ω (for example, $\omega(a) = -a^t$).
3. G is graded as $G = \oplus_{n\in\mathbf{Z}} G_n$ with G_n finite dimensional and with ω acting as -1 on the "Cartan subalgebra" G_0. (For example, we could put the basis element $e_{i,j}$ of $M_n(\mathbf{R})$ in G_{i-j}.)
4. $(a, \omega(a)) > 0$ if $g \in G_n$, $g \neq 0$.

Conversely any Lie algebra satisfying the conditions above is essentially a sum of finite dimensional and affine Lie algebras. Generalized Kac-Moody algebras are defined by the same conditions with one small change: we replace condition 4 by

4'. $(a, \omega(a)) > 0$ if $g \in G_n$, $g \neq 0$ and $n \neq 0$.

This has the effect of allowing an enormous number of new examples, such as all Kac-Moody algebras and the Heisenberg Lie algebra (which behaves like a sort of degenerate affine Lie algebra). Generalized Kac-Moody algebras have many of the properties of finite dimensional semisimple Lie algebras, and in particular they have an analogue of the Weyl character formula for some of their representations, and an analogue of the Weyl denominator formula. An example of the Weyl-Kac denominator formula for the algebra $G = SL_2[z, z^{-1}]$ is

$$\prod_{n>0}(1-q^{2n})(1-q^{2n-1}z)(1-q^{2n-1}z^{-1}) = \sum_{n\in\mathbf{Z}}(-1)^n q^{n^2} z^n.$$

This is the Jacobi triple product identity, and is also the Macdonald identity for the affine Lie root system corresponding to A_1.

Dyson described Macdonald's discovery of the Macdonald identities in [D]. Dyson found identities for $\eta(\tau)^m = q^{m/24}\prod_{n>0}(1-q^n)^m$ for the following values of m:

$$3, 8, 10, 14, 15, 21, 24, 26, 28, \ldots$$

and wondered where this strange sequence of numbers came from. (The case $m = 3$ is just the Jacobi triple product identity with $z = 1$.) Macdonald found his identities corresponding to affine root systems, which gave an explanation for the sequence above: with one exception, the numbers are the dimensions of simple finite dimensional complex Lie algebras. The exception is the number 26 (found by Atkin), which as far as I know has not been explained in terms of Lie algebras. It seems possible that it is somehow related to the fake monster Lie algebra and the special dimension 26 in string theory.

Next we give a quick explanation of "completely replicable" functions. A function is called completely replicable if its coefficients satisfy certain relations. As an example of a completely replicable function, we will look at the elliptic modular function $j(\tau) - 744 = \sum c(n)q^n$. This satisfies the identity

$$j(\sigma) - j(\tau) = p^{-1}\prod_{\substack{m>0\\ n\in\mathbf{Z}}}(1-p^m q^n)^{c(mn)}$$

where $p = e^{2\pi i\sigma}$, $q = e^{2\pi i\tau}$. (This formula was proved independently in the 80's by Koike, Norton, and Zagier, none of whom seem to have published their proofs.) Comparing coefficients of $p^m q^n$ on both sides gives many relations between the coefficients of j whenever we have a solution of $m_1 n_1 = m_2 n_2$ in positive integers, which are more or less the relations needed to show that j is completely replicable. For example, from the relation $2 \times 2 = 1 \times 4$ we get the relation

$$c(4) = c(3) + \frac{c(1)^2 - c(1)}{2}$$

or equivalently

$$20245856256 = 864299970 + \frac{196884^2 - 196884}{2}.$$

In the rest of this paper we will discuss various extensions of the original moonshine conjectures, some of which are still unproved. The first are Norton's "generalized moonshine" conjectures [N]. If we look at the Hauptmodul $T_{2A}(\tau) = q^{-1} + 4372q + \dots$ we notice that one of the coefficients is almost the same as the dimension 4371 of the smallest nontrivial irreducible representation of the baby monster simple group, and the centralizer of an element of type 2A in the monster is a double cover of the baby monster. Similar things happen for other elements of the monster, suggesting that for each element g of the monster there should be some sort of graded moonshine module $V_g = \oplus_n V_{g,n}$ acted on by a central extension of the centralizer $Z_M(g)$. In particular we would get series $T_{g,h}(\tau) = \sum_n Tr(h|V_{g,n})q^n$ satisfying certain conditions. Some progress has been made on this by Dong, Li, and Mason [D-L-M], who proved the generalized moonshine conjectures in the case when g and h generate a cyclic group by reducing to the case when $g = 1$ (the ordinary moonshine conjectures). G. Höhn has made some progress in the harder case when g and h do not generate a cyclic group by constructing the required modules for the baby monster (when g is of type $2A$). It seems likely that his methods would also work for the Fischer group Fi_{24}, but it is not clear how to go further than this. There might be some relation to elliptic cohomology (see [Hi]for more discussion of this), as this also involves pairs of commuting elements in a finite group and modular forms.

The space V_g mentioned above does not always have an invariant vertex algebra structure on it. Ryba discovered that a vertex algebra structure sometimes magically reappears when we reduce V_g modulo the prime p equal to the order of g. In fact V_g/pV_g can often be described as the Tate cohomology group $\hat{H}^0(g, V)$ for a suitable integral form V of the monster vertex algebra. This gives natural examples of vertex algebras over finite fields which do not lift naturally to characteristic 0. (Note that most books and papers on vertex algebras make the assumption that we work over a field of characteristic 0; this assumption is often unnecessary and excludes many interesting examples such as the one above.)

We will finish by describing some more of McKay's observations about the monster, which so far are completely unexplained. The monster has 9 conjugacy classes of elements that can be written as the product of two involutions of type $2A$, and their orders are 1, 2, 3, 4, 5, 6, 2, 3, 4. McKay pointed out that these are exactly the numbers appearing on an affine E_8 Dynkin diagram giving the linear relation between the simple roots. They are also the degrees of the irreducible representations of the binary icosahedral group. A similar thing happens for the baby monster: this time there are 5 classes of elements that

are the product of two involutions of type $2A$ and their orders are 2, 4, 3, 2, 1. (This is connected with the fact that the baby monster is a "3,4-transposition group".) These are the numbers on an affine F_4 Dynkin diagram, and if we take the "double cover" of an F_4 Dynkin diagram we get an E_7 Dynkin diagram. The number on an E_7 Dynkin diagram are 1, 1, 2, 2, 3, 3, 4, 2 which are the dimensions of the irreducible representations of the binary octahedral group. The double cover of the baby monster is the centralizer of an element of order 2 in the monster. Finally a similar thing happens for $Fi_{24}.2$: this time there are 3 classes of elements that are the product of two involutions of type $2A$ and their orders are 2, 3, 1. (This is connected with the fact that $F_{24}.2$ is a "3-transposition group".) These are the numbers on an affine G_2 Dynkin diagram, and if we take the "triple cover" of an G_2 Dynkin diagram we get an E_6 Dynkin diagram. The number on an E_6 Dynkin diagram are 1, 1, 1, 2, 2, 2, 3, which are the dimensions of the irreducible representations of the binary tetrahedral group. The triple cover of $Fi_{24}.2$ is the centralizer of an element of order 3 in the monster.

The connection between Dynkin diagrams and 3-dimensional rotation groups is well understood (and is called the McKay correspondence), but there is no known explanation for the connection with the monster.

References.

[B94] R. E. Borcherds, Simple groups and string theory, First European congress of mathematics, Paris July 1992, Ed. A. Joseph and others,vol. 1 p. 411-421, Birkhauser 1994.

[B98] R. E. Borcherds, Automorphic forms and Lie algebras. Current developments in mathematics 1996, to be published by international press, 1998. Also available from www.dpmms.cam.ac.uk/˜reb.

[C-G] C. J. Cummins, T. Gannon, Modular equations and the genus zero property of moonshine functions. Invent. Math. 129 (1997), no. 3, 413–443.

[C-N] J. H. Conway, S. Norton, Monstrous moonshine, Bull. London. Math. Soc. 11 (1979) 308-339.

[D] F. J. Dyson, Missed opportunities. Bull. Amer. Math. Soc. 78 (1972), 635–652.

[D-L-M] C. Dong, H. Li, G. Mason, Modular invariance of trace functions in orbifold theory, preprint q-alg/9703016.

[F-L-M] I. B. Frenkel, J. Lepowsky, A. Meurman, Vertex operator algebras and the monster, Academic press 1988. (Also see the announcement A natural representation of the Fischer-Griess monster with the modular function J as character, Proc. Natl. Acad. Sci. USA 81 (1984), 3256-3260.)

[Ge] R. W. Gebert, Introduction to vertex algebras, Borcherds algebras and the monster Lie algebra, Internat. J. Modern Physics A8 (1993) no 31, 5441-5503.

[Go] The Work of R.E. Borcherds, this volume, preprint math/9808136.

[G-T] P. Goddard and C. B. Thorn, Compatibility of the dual Pomeron with unitarity and the absence of ghosts in the dual resonance model, Phys. Lett., B 40, No. 2 (1972), 235-238.

[G] R. L. Griess, The friendly giant. Invent. Math. 69 (1982), no. 1, 1–102.

[Hi] F. Hirzebruch, T. Berger, R. Jung, Manifolds and modular forms. Aspects of Mathematics, E20. Friedr. Vieweg & Sohn, Braunschweig, 1992. ISBN: 3-528-06414-5

[H] G. Höhn, Selbstduale Vertexoperatorsuperalgebren und das Babymonster. [Self-dual vertex-operator superalgebras and the Baby Monster] Dissertation, Rheinische Friedrich-Wilhelms-Universität Bonn, Bonn, 1995. Bonner Mathematische Schriften 286. Universität Bonn, Mathematisches Institut, Bonn, 1996.

[J] E. Jurisich, Generalized Kac-Moody Lie algebras, free Lie algebras and the structure of the Monster Lie algebra. J. Pure Appl. Algebra 126 (1998), no. 1-3, 233–266.

[K90] V. G. Kac, "Infinite dimensional Lie algebras", third edition, Cambridge University Press, 1990. (The first and second editions (Birkhauser, Basel, 1983, and C.U.P., 1985) do not contain the material on generalized Kac-Moody algebras.)

[Ko] M. Koike, On Replication Formula and Hecke Operators, Nagoya University preprint.

[LZ] B. Lian, G. Zuckerman, New perspectives on the BRST-algebraic structure in string theory, hepth/921107, Communications in Mathematical Physics 154, (1993) 613-64, and Moonshine cohomology, q-alg/ 950101 Finite groups and Vertex Operator Algebras, RIMS publication (1995) 87-11.

[M] Y. Martin, On modular invariance of completely replicable functions. Moonshine, the Monster, and related topics (South Hadley, MA, 1994), 263–286, Contemp. Math., 193, Amer. Math. Soc., Providence, RI, 1996.

[N] S. P. Norton, Appendix to G. Mason's paper in The Arcata Conference on Representations of Finite Groups (Arcata, Calif., 1986), 181–210, Proc. Sympos. Pure Math., 47, Part 1, Amer. Math. Soc., Providence, RI, 1987.

Commun. Math. Phys. 119, 221–241 (1988)

Communications in Mathematical Physics
© Springer-Verlag 1988

Beauty and the Beast: Superconformal Symmetry in a Monster Module

L. Dixon[1,*], P. Ginsparg[2,**] and J. Harvey[3,***]

[1,3] Physics Department, Princeton University, Princeton, NJ 08544, USA

[2] Lyman Laboratory of Physics, Harvard University, Cambridge, MA 02138, USA

Abstract. Frenkel, Lepowsky, and Meurman have constructed a representation of the largest sporadic simple finite group, the Fischer–Griess monster, as the automorphism group of the operator product algebra of a conformal field theory with central charge $c = 24$. In string terminology, their construction corresponds to compactification on a $\mathbf{Z}_2$ asymmetric orbifold constructed from the torus $\mathbf{R}^{24}/\Lambda$, where Λ is the Leech lattice. In this note we point out that their construction naturally embodies as well a larger algebraic structure, namely a super-Virasoro algebra with central charge $\hat{c} = 16$, with the supersymmetry generator constructed in terms of bosonic twist fields.

1. Introduction

A fundamental example of the techniques of conformal field theory arises in the construction of a moonshine module for the Fischer–Griess monster (the largest sporadic simple finite group) by Frenkel, Lepowsky, and Meurman (FLM) [1]. While this construction explains some of the puzzling connections between the representation theory of the monster and modular forms (referred to as moonshine in [2]), many aspects of this connection remain obscure.

In this note we point out a further feature of the construction, namely the existence of an underlying superconformal algebra in which the fermionic component of the super stress-energy tensor is constructed in terms of bosonic twist fields. In Sects. 2 and 3, we will first present a brief summary of the Virasoro and super-Virasoro algebras and their representations in terms of operator product expansions. In Sects. 4 and 5, we then present a brief synopsis of the FLM construction of the monster in the language of conformal field theory. We have tried to make these sections particularly accessible to string theorists, who may wish to satisfy their cultural curiosity in the "monstrous game." In Sect. 6 we

* (ljd@pupthy princeton.edu or ljd@pucc.bitnet)

** (ginsparg@huhepl.hepnet, ginsparg@huhepl.bitnet, or ginsparg@huhepl.harvard.edu)

*** (jah@pupthy.princeton.edu or jah@pucc.bitnet)

demonstrate the existence of a superconformal algebra in the construction, and show how the states in the theory are organized by the superconformal symmetry. In Sect. 7 we give a brief overview of the genus zero function field problem, an important open problem that links seemingly unrelated fields of mathematics. Finally in Sect. 8 we close with a discussion of issues related to those of this paper.

Our essential observation is that the $\mathbf{Z}_2$ twist operator for 24 bosonic fields has dimension $24\cdot(1/16) = 3/2$, where $1/16$ is the dimension of a single $\mathbf{Z}_2$ bosonic twist field. These dimension $3/2$ operators indeed prove to have all the correct properties to serve as supersymmetry generators for holomorphic conformal field theories based on $\mathbf{R}^{24}/\Lambda$ for any Niemeier lattice Λ, and in particular for the case $\Lambda = \Lambda_{\text{Leech}}$ used in the FLM construction. We hope that the additional algebraic structure revealed here will be useful in resolving some of the lingering mysterious features of the monster. We suspect also that the search for a unified understanding of all finite groups based on lattices, vertex operators, and loop algebras may be a mathematical arena where "physics" ideas such as supersymmetry will prove crucial.

2. Conformal Field Theory

In two-dimensional conformal field theory, analytic conformal transformations of the form $z \to z + \varepsilon z^{n+1}$ are generated by the moments

$$L_n = \oint \frac{dz}{2\pi i} z^{n+1} T(z) \tag{2.1}$$

of the stress tensor $T(z)$. The L_n's satisfy the commutation relations

$$[L_m, L_n] = (m-n)L_{m+n} + \frac{c}{12}(m^3 - m)\delta_{m+n,0} \tag{2.2}$$

of the Virasoro algebra, where c is a constant ($[L_n, c] = 0$) known as the central charge of the algebra. A general conformal field theory will involve as well anti-analytic conformal transformations generated by $\bar{L}_n$'s. The commutation relations (2.2) are equivalent to the operator product expansion (OPE)

$$T(z_1)T(z_2) \sim \frac{c/2}{(z_1 - z_2)^4} + \frac{2}{(z_1 - z_2)^2} T(z_2) + \frac{1}{(z_1 - z_2)} \partial_2 T(z_2) + \cdots, \tag{2.3}$$

where we shall systematically suppress the non-singular terms in what follows. The equivalence between (2.2) and (2.3) follows from standard contour deformation arguments. (For introductions to standard conformal field theory techniques, see e.g. [3].)

Fields that satisfy the OPE

$$T(z_1)\phi(z_2) \sim \frac{h}{(z_1 - z_2)^2} \phi(z_2) + \frac{1}{(z_1 - z_2)} \partial_2 \phi(z_2) + \cdots \tag{2.4}$$

are termed primary fields of dimension, or conformal weight, h. Each primary field is associated with a highest weight state $|h\rangle = \phi(0)|0\rangle$, where $|0\rangle$ is the

vacuum state (associated with the identity operator). Equation (2.4) implies that $L_0|h\rangle = h|h\rangle$ and $L_n|h\rangle = 0$, $n > 0$. Applying the "creation" operators $L_n, n < 0$, to any highest weight $|h\rangle$ gives a (reducible) representation of the Virasoro algebra called a Verma Module, which we denote $\mathbf{M}_{h,c}$. This module has an integer grading $\mathbf{M}_{h,c} = \oplus_n \mathbf{M}^n_{h,c}$, where n is the eigenvalue of $L_0 - h$, i.e.

$$\mathbf{M}^n_{h,c} = \{v \in \mathbf{M}_{h,c} | L_0 v = (h+n)v\}. \tag{2.5}$$

Its character is given by

$$\operatorname{ch} \mathbf{M}_{h,c}(q) \equiv q^{h-c/24} \sum_{n=0}^{\infty} (\dim \mathbf{M}^n_{h,c}) q^n = \frac{q^{h-c/24}}{\prod_{n=1}^{\infty} (1-q^n)}. \tag{2.6}$$

From a physical point of view, the factor of $q^{-c/24}$ introduced in (2.6) arises as a vacuum energy; from a mathematical point of view, it improves the modular transformation properties of the partition function (to be discussed below).

Unitarity of representations of the Virasoro algebra in general restricts the values of h and c [4]. In many cases the Verma module contains null states, which means that it provides a reducible representation of the Virasoro algebra, and one must factor out the null states to obtain an irreducible module, denoted by $\mathbf{L}_{h,c}$. When $c > 1$, however, all representations with $h \geqq 0$ are unitary, and all Verma modules with $h > 0$ are irreducible, so we have

$$\begin{aligned} \operatorname{ch} \mathbf{L}_{h,c}(q) &= \operatorname{ch} \mathbf{M}_{h,c}(q), \quad c > 1, \quad h > 0, \\ \operatorname{ch} \mathbf{L}_{0,c}(q) &= \frac{q^{-c/24}}{\prod_{n=2}^{\infty} (1-q^n)}. \end{aligned} \tag{2.7}$$

In a general conformal field theory, highest weight states are labelled by their conformal weights $(h, \bar{h})$, i.e. their eigenvalues under L_0 and $\bar{L}_0$ respectively. The genus one partition function of the theory is given by

$$Z_c(q, \bar{q}) = \sum_{h,\bar{h}} N_{h,\bar{h}} \operatorname{ch} \mathbf{L}_{h,c}(q) \operatorname{ch} \mathbf{L}_{\bar{h},c}(\bar{q}), \tag{2.8}$$

where $N_{h,\bar{h}}$ is an integer equal to the number of primary fields with conformal weight $(h, \bar{h})$. Moreover if we write $q = \exp(2\pi i \tau)$, where τ is the modular parameter for a torus, then $Z_c(q, \bar{q})$ should be invariant under modular transformations

$$\tau \to \frac{a\tau + b}{c\tau + d}, \quad \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in SL(2, \mathbf{Z}). \tag{2.9}$$

For $c > 1$ this requires that the number of primary fields be infinite [5]. In certain cases, $Z_c(q, \bar{q})$ will factorize into holomorphic and anti-holomorphic parts, both of which are modular covariant, and the anti-holomorphic part may be ignored. (Even in this case, however, generalization of the partition function to higher genus Riemann surfaces may require inclusion of both holomorphic and anti-holomorphic parts [6].) In what follows we shall concentrate exclusively on the case of purely holomorphic conformal field theories.

3. Superconformal Field Theory

Supersymmetric extensions of the Virasoro algebra are obtained by generalizing conformal transformations to superconformal transformations of supercoordinates $\mathbf{z}=(z,\theta)$, where θ is an anticommuting coordinate ($\theta^2=0$). Superconformal transformations are generated by the moments of a super stress-energy tensor. If there is only a single anti-commuting coordinate ($N=1$ supersymmetry) then the super stress-energy tensor $\mathbf{T}(\mathbf{z})=T_F(z)+\theta T_B(z)$ has components that satisfy the operator products

$$\begin{aligned}
T_B(z_1)T_B(z_2) &\sim \frac{3\hat{c}/4}{(z_1-z_2)^4}+\frac{2}{(z_1-z_2)^2}T_B(z_2)+\frac{1}{(z_1-z_2)}\partial_2 T_B(z_2),\\
T_B(z_1)T_F(z_2) &\sim \frac{3/2}{(z_1-z_2)^2}T_F(z_2)+\frac{1}{(z_1-z_2)}\partial_2 T_F(z_2),\\
T_F(z_1)T_F(z_2) &\sim \frac{\hat{c}/4}{(z_1-z_2)^3}+\frac{1/2}{(z_1-z_2)}T_B(z_2),
\end{aligned} \tag{3.1}$$

where $\hat{c}=\frac{2}{3}c$. The conventional normalization is such that the stress-energy tensor for a single free bosonic field $x(z)$ has central charge $c=1$ in (2.3), and that for a single free superfield $x(z)+\theta\psi(z)$ has central charge $\hat{c}=1$ in (3.1). The second equation in (3.1) is the statement that T_F is a primary field of dimension 3/2.

In terms of the moments L_n of T_B, and the moments

$$G_n=\oint\frac{dz}{2\pi i}z^{n+1/2}2T_F(z) \tag{3.2}$$

of T_F, the OPE's (3.1) are equivalent to the (anti-)commutation relations

$$\begin{aligned}
[L_m,L_n] &= (m-n)L_{m+n}+\frac{\hat{c}}{8}(m^3-m)\delta_{m+n,0},\\
[L_m,G_n] &= \left(\frac{m}{2}-n\right)G_{m+n},\\
\{G_m,G_n\} &= 2L_{m+n}+\frac{\hat{c}}{2}(m^2-\tfrac{1}{4})\delta_{m+n,0}.
\end{aligned} \tag{3.3}$$

For integer moding ($n\in\mathbf{Z}$) of G_n, the supersymmetric extension of the Virasoro algebra is termed the Ramond (R) algebra; for half-integer moding ($n\in\mathbf{Z}+\frac{1}{2}$), it is termed the Neveu–Schwarz (NS) algebra. Primary fields are again associated with highest weight states $|h\rangle$, satisfying $L_n|h\rangle=G_n|h\rangle=0, n>0$, and $L_0|h\rangle=h|h\rangle$. Note that (3.3) requires that a highest weight state in the Ramond sector have eigenvalue $h-\hat{c}/16$ under G_0^2. For $\hat{c}>1$, the only restrictions imposed by unitarity are $h\geqq 0$ (NS), and $h\geqq\hat{c}/16$ (R), and the Verma modules again provide irreducible representations unless the latter inequalities are saturated.

In order to obtain a modular invariant partition function, it is necessary to include representations of both the R and NS algebras together with a projection

onto a subset of the highest weight states [7]. This is accomplished by introducing an operator $\hat{\Gamma}$ that commutes with T_B and $\bar{T}_B$, anticommutes with T_F and $\bar{T}_F$, and satisfies $\hat{\Gamma}^2 = 1$. The modular invariant partition function then takes the form

$$Z(q, \bar{q}) = (q\bar{q})^{-\hat{c}/16}(\mathrm{Tr}_{\mathrm{NS}} P q^{L_0} \bar{q}^{\bar{L}_0} + \mathrm{Tr}_{\mathrm{R}} P q^{L_0} \bar{q}^{\bar{L}_0}), \tag{3.4}$$

where the projection $P = (1 + \hat{\Gamma})/2$. In the Ramond algebra, G_0 commutes with L_0 and satisfies $G_0^2 = L_0 - \hat{c}/16$. Thus if $h > \hat{c}/16$, highest weight states of the Virasoro algebra are paired by G_0 with states having opposite eigenvalue under $\hat{\Gamma}$. Only the Ramond ground states, with $h = \hat{c}/16$, are not so paired since they may be annihilated by G_0. It follows that the contribution

$$\pm (q\bar{q})^{-\hat{c}/16} \mathrm{Tr}_{\mathrm{R}} \hat{\Gamma} q^{L_0} \bar{q}^{\bar{L}_0} \tag{3.5}$$

to the partition function is always a constant (analogous to the Witten index [8]). The value of this constant is not determined by genus-one modular invariance, but rather by the requirement of a consistent operator interpretation of the terms in (3.4). If a nonzero value of the constant is permitted, the two sign choices in (3.5) generate inequivalent modular invariant theories [9, 10], each consistent, and distinguished from one another by an opposite choice of projection in the Ramond sector.

It is not always obvious whether a given conformal field theory has a superconformal symmetry. One problem is that the state with conformal weight $(\frac{3}{2}, 0)$ created by T_F itself cannot survive in a modular invariant theory, since invariance under $\tau \to \tau + 1$ allows only integer spin $s = h - \bar{h}$ for all physical states. Before projection on $\hat{\Gamma} = +1$ states however, the $\hat{\Gamma} = -1$ states of the theory will include in general states with half-integer spin. Projection by P removes all these states and leaves a modular invariant local theory. To identify a superconformal algebra, one must therefore look beyond the physical states and construct a non-local $\mathbf{Z}_2$ cover of the theory containing a weight $(\frac{3}{2}, 0)$ primary field that satisfies the OPE (3.1). The $\mathbf{Z}_2$ cover will consist of the set of all fields in the original conformal field theory, plus T_F, plus any fields needed to close the operator algebra of T_F with the original fields. *All* OPE's in the covering theory are either local (i.e. have integer power singularities), or have at worst square-root type singularities. Finally, there is a projection onto $\mathbf{Z}_2$ invariant operators of some $\mathbf{Z}_2$ symmetry that yields the original local conformal field theory. (A general method for constructing this non-local $\mathbf{Z}_2$ cover will be outlined at the end of Sect. 6. It makes use of the $\mathbf{Z}_2$ symmetry-possessed by any superconformal field theory-that acts as $+1$ and -1 on Neveu–Schwarz and Ramond sectors respectively.)

Conversely, given a local theory with a non-local $\mathbf{Z}_2$ cover that contains a dimension $(\frac{3}{2}, 0)$ field satisfying (3.1), one can always organize the states into representations of a superconformal algebra with the $(\frac{3}{2}, 0)$ field playing the role of T_F. Fields in the Ramond sector will be those with square-root singularities in their OPE with T_F, while those in the NS sector will have integer power singularities. The intent here is to show that the holomorphic conformal field theory used in the FLM construction of the monster possesses precisely this structure.

4. Lattices and Holomorphic Conformal Field Theory

Given an even, self-dual, integer lattice Λ of dimension r, we can always define a holomorphic modular-covariant conformal field theory with central charge $c=r$. First we introduce r free holomorphic fields $x^i(z), i=1,\ldots,r$, that we view as coordinates on the torus $\mathbf{T}^r=\mathbf{R}^r/\Lambda$, i.e. $x\equiv x+2\pi\beta, \beta\in\Lambda$. The operators in their mode expansions[1]

$$x^i(z)=q^i-ip^i\ln z+i\sum_{n\neq 0}\frac{\alpha_n^i}{n}z^{-n} \tag{4.1}$$

satisfy the commutation relations

$$[q^i,p^j]=-i\delta^{ij},\quad [\alpha_n^i,\alpha_m^j]=n\delta^{ij}\delta_{n+m,0}. \tag{4.2}$$

For every lattice vector $\beta\in\Lambda$, we can construct a Fock space F_β. It is built on the highest weight state $|\beta\rangle$, which satisfies $\alpha_n|\beta\rangle=0$ for $n>0$ and $p^i|\beta\rangle=\beta^i|\beta\rangle$ (β^i are the components of β in an orthonormal frame). States of these Fock spaces are constructed by acting on $|\beta\rangle$ with the creation operators $\alpha_{-n}, n>0$.

A holomorphic conformal field theory may be defined via a field theory action as follows. We begin with the action

$$S=\frac{1}{2\pi}\int\partial X^i\bar\partial X^i+B_{ij}\partial X^i\bar\partial X^j=\frac{1}{2\pi}\int g_{ab}\partial X^a\bar\partial X^b+b_{ab}\partial X^a\bar\partial X^b, \tag{4.3}$$

where the fields $X^i(z,\bar z)\in\mathbf{R}^r/(\Lambda/2)$, i.e. $X^i\equiv X^i+2\pi(\beta^i/2)$, $\beta\in\Lambda$. We take the even, self-dual, lattice Λ of dimension r to be generated by basis vectors $e_a, a=1,\ldots,r$, so that $\beta=n^a e_a, n^a\in\mathbf{Z}$, and define a metric $g_{ab}=e_a\cdot e_b$. The coordinates X^a in (4.3) are defined by $X^i=(e_a)^iX^a$, and are thus periodic up to multiples of $\pi, X^a\equiv X^a+2\pi(n^a/2)$. Note that the vectors $e^{*a}=g^{ab}e_b$, defined in terms of the inverse metric $g^{ab}=g_{ab}^{-1}$, satisfy $e^{*a}\cdot e_b=\delta^a_b$, and by self-duality are consequently also a basis of Λ. Finally $b_{ab}=e_a\cdot B\cdot e_b=B_{ij}(e_a)^i(e_b)^j$ are the components of the (constant) antisymmetric B-field referred to lattice frames.

The partition function for the theory (4.3) is [11]

$$Z(q,\bar q)=(q\bar q)^{-r/24}\,\mathrm{Tr}\,q^{L_0}\bar q^{\bar L_0}=\frac{1}{(\eta\bar\eta)^r}\sum_{w\in\Lambda/2,p\in 2\Lambda}q^{(1/2)p_L^2}\bar q^{(1/2)p_R^2}, \tag{4.4}$$

where $\eta(q)=q^{1/24}\prod_{n=1}^{\infty}(1-q^n)$ is the Dedekind eta function, and $p_{L,R}=\frac{1}{2}p-Bw\pm w=m_a e^{*a}-\frac{1}{2}(b_{ab}\mp g_{ab})e^{*a}n^b$. In general the summations in (4.4) are not consistent with a splitting between holomorphic and anti-holomorphic Hilbert spaces. But if the B-field is chosen to satisfy

$$e_a\cdot B\cdot e_b=e_a\cdot e_b\mod 2, \tag{4.5}$$

(e.g. take $b_{ab}=\pm g_{ab}, a\leqslant b$), the summation in (4.4) is easily seen to decompose into independent summations over $p_L\in\Lambda$ and $p_R\in\Lambda$. Equation (4.4) then takes the

[1] Our normalization conventions are such that the operator products satisfy $\partial x^i(z)\partial x^j(w)\sim-\delta^{ij}/(z-w)^2$

factorized form

$$Z(q,\bar{q}) = \left| \frac{1}{\eta^r} \sum_{\beta \in \Lambda} q^{(1/2)\beta^2} \right|^2$$

(and we would have e.g. the factorization $N_{h,\bar{h}} = N_h N_{\bar{h}}$ in (2.8)).

In the case at hand, one can show as well that the Hilbert space decomposes into a product of holomorphic and anti-holomorphic pieces and one can therefore impose a consistent projection onto the purely holomorphic part. This is the theory that we shall consider in what follows, i.e. we shall retain only the dependence on the coordinate $x(z)$, where $X(z,\bar{z}) = \frac{1}{2}(x(z) + x(\bar{z}))$. (This defines what we mean by holomorphic conformal field theory on $\mathbf{R}^r/\Lambda$, although strictly speaking we should probably call it something like $\mathbf{R}^r/(2\pi\Lambda)^{1/2}$.)

The stress-energy tensor of the theory is defined as

$$T(z) = -\tfrac{1}{2} :\partial_z x \cdot \partial_z x:, \tag{4.6}$$

(where the double dots indicate normal ordering), with the corresponding Virasoro generators given by

$$L_n = \frac{1}{2} \sum_{m=-\infty}^{m=+\infty} \alpha_{n-m} \cdot \alpha_m, \quad n \neq 0, \tag{4.7}$$

where $\alpha_0^i = p^i$. To define L_0 we must resolve the operator ordering ambiguity that arises due to the non-commutativity of α_n and α_{-n}. The definition

$$L_0 = \tfrac{1}{2}p^2 + \sum_{n=1}^{\infty} \alpha_{-n} \cdot \alpha_n, \tag{4.8}$$

with the vacuum normal ordering constant instead included in the prefactor to the partition function, satisfies the commutation relations (2.2).

Primary fields are given by particular products of the dimension one $U(1)$ currents $j^i(z) = i\partial_z x^i$ and the dimension $\beta^2/2$ exponentials

$$V_\beta(z) =: e^{i\beta \cdot x(z)}:, \quad \beta \in \Lambda. \tag{4.9}$$

We see that the vertex operators V_β are well defined under $x \to x + 2\pi\alpha$, $\alpha \in \Lambda$, since then $\beta \cdot \alpha \in \mathbf{Z}$. Strictly speaking, (4.9) is incorrect because the OPE of these fields is not associative. Associativity of the OPE is restored by constructing [12] a projective representation $\bar{\Lambda}$ of the lattice Λ with two-cocyle $\varepsilon(\alpha,\beta)$, and including on the right-hand side of (4.9) a factor $c(\alpha)$ that satisfies $c(\alpha)c(\beta) = \varepsilon(\alpha,\beta)c(\alpha+\beta)$. This projective representation for the Leech lattice is described in detail by FLM. In terms of the antisymmetric B field mentioned above, we note that we can represent $\varepsilon(\alpha,\beta) = e^{i\pi\alpha\cdot B\cdot\beta/2}$. It then follows from (4.5) that $\varepsilon(\alpha,\beta)\varepsilon^{-1}(\beta,\alpha) = (-1)^{\alpha\cdot B\cdot\beta} = (-1)^{\alpha\cdot\beta}$, as required for associativity of the OPE's of the $c(\alpha):e^{i\alpha\cdot x(z)}:$'s. (In terms of the momentum operator p, the $c(\alpha)$'s themselves can be represented in this case as $c(\alpha) = e^{i\pi\alpha\cdot B\cdot p/2}$.)

The partition function for the theory is given by

$$Z(q) = q^{-c/24} \operatorname{Tr} q^{L_0} = \frac{\Theta_\Lambda(q)}{\eta^r(q)}, \tag{4.10}$$

where $\Theta_\Lambda(q) = \sum_{\beta\in\Lambda} q^{(1/2)\beta^2}$ is the theta function of the lattice Λ. For c a multiple of 24, $Z(q)$ will be modular invariant; for $c = 8, 16 \bmod 24$, on the other hand, $Z(q)$ will be modular covariant, i.e. a weight-zero modular function with non-trivial multiplier (in other words for the cases above it picks up a phase $e^{2\pi ic/24}$ under the modular transformation $\tau \to \tau + 1$). An example of a holomorphic conformal field theory arises for the choice $\Lambda = \Gamma_8$, where Γ_8 is the root lattice of E_8. The holomorphic partition function is then the character of the (basic) representation of the level one affine Kac–Moody algebra $\hat{E}_8$. It is thus given by $\Theta_{\Gamma_8}/\eta^8 = j^{1/3}$, where the famous weight-zero modular function j is given by $j(q) = q^{-1} + 744 + 196884q + \cdots$.

In what follows we shall ultimately be interested in the case $\Lambda = \Lambda_{\text{Leech}}$, where the Leech lattice Λ_{Leech} is the unique dimension 24, even, self-dual lattice with $\Lambda_2 = \emptyset$ (Λ_n is the set of points in Λ with length-squared n). For this case, the holomorphic conformal field theory constructed by the above procedure has partition function

$$Z_{\text{Leech}}(q) = \frac{\Theta_{\Lambda_{\text{Leech}}}}{\eta^{24}} = \frac{1}{q} + 24 + 196884q + \cdots = J(q) + 24, \tag{4.11}$$

with $J(q) = j(q) - 744$ (i.e. J has constant term set equal to zero). For $c = 24$ we see that the partition function is indeed modular invariant. Part of monstrous moonshine is the fact that the terms in the expansion (4.11), apart from the constant term, are simple sums of the dimensions of irreducible representations of the monster. In the next section we shall review how the FLM construction removes the unwanted constant term.

Given a modular invariant conformal field theory with a discrete symmetry group G, it is often possible to construct a new conformal field theory by the procedure of "modding out" or "twisting" by G [13]. This procedure can be applied to holomorphic conformal field theories based on even, self-dual lattices Λ, and modular invariance puts stringent restrictions on G. In the simplest case G will be a $\mathbf{Z}_n$ automorphism of Λ, generated by the $SO(r)$ rotation (in a diagonal basis)

$$g = \text{diag}(e^{2\pi i\xi_1/n}, \ldots, e^{2\pi i\xi_l/n}; e^{-2\pi i\xi_1/n}, \ldots, e^{-2\pi i\xi_l/n}), \tag{4.12}$$

where $l = r/2$. Then the restriction on G is [13, 14]

$$\sum_{i=1}^{l} \xi_i^2 = 0 \begin{cases} \bmod n & n \text{ odd} \\ \bmod 2n & n \text{ even} \end{cases}. \tag{4.13}$$

For n even, there is an additional restriction particular to asymmetric bosonic orbifolds [15, 16],

$$\beta \cdot g^{n/2} \beta = 0 \bmod 2, \quad \text{for all } \beta \in \Lambda. \tag{4.14}$$

Since even, self-dual, lattices exist only in $8k$ dimensions, a twist by an involutive automorphism g defined to take $x \to -x$, and hence lattice vectors $\beta \to -\beta$, will always satisfy (4.13)–(4.14). This is because $-\beta^2 = 0 \bmod 2$, and we have $n = 2$ and all the $\xi_i = 1$ in (4.12), giving $\sum \xi_i^2 = l = 4k$. In the same way that we can view the conformal field theory determined by Λ as corresponding to string compactification

on the torus $\mathbf{T}^r = \mathbf{R}^r/\Lambda$, we can view the twisting procedure as corresponding to string compactification on the orbifold $\mathbf{T}^r/\mathbf{Z}_2$. Note that the $\mathbf{Z}_2$ twists that we consider here have the simplifying feature that they leave no sublattice of Λ invariant, so the fixed points are classified by $\Lambda/2\Lambda$.

To implement the twisting procedure for $\mathbf{Z}_2$ twists, we first construct the untwisted Hilbert space $H_{(+)}$ as a direct product of the Fock spaces F_β. $H_{(+)}$ has a $\mathbf{Z}_2$ grading induced by the action of the generator g of $G = \mathbf{Z}_2$, and we define

$$H^{\pm}_{(+)} = \{v \in H_{(+)} | gv = \pm v\}. \tag{4.15}$$

To construct a twisted Hilbert space $H_{(-)}$, we take fields $x^i(z)$ satisfying boundary conditions $x^i(e^{2\pi i}z) = -x^i(z) \pmod{\Lambda}$ with mode expansion

$$x^i(z) = \tilde{q}^i + i \sum_{n \in Z + 1/2} \frac{\alpha^i_n}{n} z^{-n}, \tag{4.16}$$

where α^i_n obey the same commutation relations (4.2) as before.

For a $\mathbf{Z}_2$ twist acting on both holomorphic and anti-holomorphic parts of $X(z, \bar{z}) = \frac{1}{2}(x(z) + x(\bar{z}))$, $\tilde{q}^i$ would take values in the 2^r dimensional fixed-point set of g, given by $\Lambda/2\Lambda$. Because the twist acts only on the holomorphic field $x(z)$, the situation is more subtle [1, 15, 16]. The two-cocycle $\varepsilon(\alpha, \beta)$ determines a $\mathbf{Z}_2$ central extension of Λ, and also a $\mathbf{Z}_2$ extension of $\Lambda/2\Lambda = (\mathbf{Z}_2)^{24}$ (viewing the lattice Λ as an abelian group). This central extension of $\Lambda/2\Lambda$ has a 2^{12} dimensional irreducible representation, acting on a space T with basis elements $|\sigma^a\rangle = \sigma^a(0)|0\rangle$, $a = 0, 1, \ldots, 2^{12} - 1$. The $|\sigma^a\rangle$ are particular linear combinations of the 2^{24} states, denoted $|\tilde{q}_b\rangle$, with definite values of $\tilde{q}$:

$$\tilde{q}|\tilde{q}_b\rangle = \tilde{q}_b|\tilde{q}_b\rangle, \quad \text{where} \quad 2\tilde{q}_b \in \Lambda/2\Lambda. \tag{4.17}$$

(The factor 2^{12} can also be regarded as the square root of the number of fixed points of the transformation $X^i(z, \bar{z}) \to -X^i(z, \bar{z})$. The 2^{24} fixed points transform as a $(2^{12}, 2^{12})$ representation of the holomorphic and anti-holomorphic symmetry groups. The projection onto the holomorphic Hilbert space removes the right 2^{12} degeneracy.) The twisted Hilbert space is then constructed as the tensor product

$$H_{(-)} = T \otimes F_{(-)}, \tag{4.18}$$

where the twisted Fock space $F_{(-)}$ is constructed by acting with the $\alpha_{-n}, n > 0$ ($n \in \mathbf{Z} + \frac{1}{2}$), on the Fock vacuum. $H_{(-)}$ also has a $\mathbf{Z}_2$ grading induced by the action of g as in (4.15), and we refer to the corresponding eigenspaces as $H^{\pm}_{(-)}$.

Using (4.6) and the mode expansion (4.16) gives for the Virasoro generators acting in the twisted Hilbert space

$$L_n = \frac{1}{2} \sum_{m \in Z + 1/2} \alpha_{n-m} \cdot \alpha_m, \quad n \neq 0,$$

$$L_0 = \sum_{m = 1/2}^{\infty} \alpha_{-m} \cdot \alpha_m + \frac{r}{16}. \tag{4.19}$$

The vacuum normal ordering constant $r/16$ for L_0 in the twisted sector (relative to the untwisted vacuum) is required in order that (4.19) satisfies the Virasoro

algebra (2.2) (as can be seen using $\frac{1}{2}\sum_{n=0}^{\infty}(n+\frac{1}{2}) = -\frac{1}{4}\zeta(-1) = \frac{1}{48} = -\frac{1}{24}+\frac{1}{16}$). (An alternate derivation of the constant will appear shortly.)

The full partition function of the twisted theory is given by

$$\begin{aligned} Z_{\mathrm{tw}}(q) &= q^{-r/24}(\mathrm{Tr}_{H^+_{(+)}} q^{L_0} + \mathrm{Tr}_{H^+_{(-)}} q^{L_0}) \\ &= q^{-r/24}(\mathrm{Tr}_{H_{(+)}} P q^{L_0} + \mathrm{Tr}_{H_{(-)}} P q^{L_0}), \end{aligned} \tag{4.20}$$

where the traces run over the indicated Hilbert spaces, and $P = (1+g)/2$. An intuitive understanding of the modular properties of the terms appearing in Z_{tw} comes from the path integral interpretation of the traces appearing in (4.20). A term of the form $\mathrm{Tr}_{H_h} g q^{L_0}$ evaluated in the Hilbert space H_h twisted by the group element h, denoted by

$$g \underset{h}{\boxed{}}, \tag{4.21}$$

can be computed as a path integral over a torus with modular parameter τ and boundary conditions twisted by g and h in the "time" and "space" directions respectively (in (4.20) we have $H_{(+)} \equiv H_1$ and $H_{(-)} \equiv H_g$). A modular transformation $\tau \to (a\tau + b)/(c\tau + d)$ acts as a global diffeomorphism of the torus and changes the boundary conditions to

$$g^a h^b \underset{g^c h^d}{\boxed{}}. \tag{4.22}$$

Z_{tw} in (4.20) is given by the sum

$$Z_{\mathrm{tw}} = \frac{1}{2}\left(1 \underset{1}{\boxed{}} + g \underset{1}{\boxed{}} + 1 \underset{g}{\boxed{}} + g \underset{g}{\boxed{}} \right), \tag{4.23}$$

which is modular invariant provided that (4.13) and (4.14) are satisfied.

In conformal field theory the $2^{r/2}$ degenerate twisted ground states belonging to T are created from the untwisted ground state by primary conformal fields known as twist fields, and denoted by σ^a. Because the σ^a change the sign of $\partial_z x$, they have the non-local OPE

$$\partial x^i(z)\sigma^a(w) \sim (z-w)^{-1/2}\tau^{ia}(w) + \cdots, \tag{4.24}$$

where τ is an excited twist field of dimension 2 that creates excited states in $H_{(-)}$. The projection onto states in $H^+_{(+)}$ and $H^+_{(-)}$ keeps a set of operators that have mutually local OPE's.

Using the OPE (4.24), one can derive the conformal dimension of the fields σ^a, which is also the normal-ordering constant for the vacuum in the twisted sector. Consider the correlation function

$$\frac{\langle 0|\sigma(\infty)\partial x^i(z)\partial x^j(w)\sigma(0)|0\rangle}{\langle 0|\sigma(\infty)\sigma(0)|0\rangle} = -\frac{\frac{1}{2}\left(\sqrt{\frac{z}{w}} + \sqrt{\frac{w}{z}}\right)\delta^{ij}}{(z-w)^2} \tag{4.25}$$

of bosonic fields in the twisted sector, where σ is a twist field for all r bosonic coordinates x^i. With the normal ordered stress-energy tensor $T(z)$ defined as $-\frac{1}{2}[\partial x(z)\cdot\partial x(w) + r/(z-w)^2]_{z\to w}$, we infer the leading operator product $T(z)\sigma(0) \sim$

$(r/16)z^{-2}\sigma(0)+\cdots$, and hence from (2.4) we have $h=r/16$. In a canonical normalization for the twist fields, their OPE's with each other are given by

$$\sigma^a(z)\sigma^b(w)\sim\delta^{ab}(z-w)^{-r/8}+\cdots. \tag{4.26}$$

For the $c=8$ E_8 holomorphic conformal field theory, it is known that the above twisting procedure gives back an isomorphic theory. (In terms of the level one $\hat{D}_8$ subcontent, the projection in the untwisted sector removes the $\hat{D}_8$ spinor representation, leaving the character for the $\hat{D}_8$ basic representation; and the projection in the twisted sector removes the $\hat{D}_8$ vector, leaving the character for the other $\hat{D}_8$ spinor [1]). In the $c=16$ case, it is known [13, 17] that this twisting procedure simply interchanges the two holomorphic conformal field theories based on the $E_8\times E_8$ and $\mathrm{Spin}(32)/\mathbf{Z}_2$ lattices. The first case for which the procedure produces new interesting holomorphic conformal field theories occurs at $c=24$, and the most interesting of these comes from modding out the conformal field theory based on the Leech lattice.

5. The Moonshine Module

In the construction of the monster, the need for the $\mathbf{Z}_2$ twisting described above is motivated by finite group theory [1]. The conformal field theory determined by the Leech lattice has a large symmetry group that acts as an automorphism group of the operator product algebra. Naively, one would expect this to be just the automorphism group $\cdot 0$ of the Leech lattice, which is also known as the Conway group. Due to the cocycle factors that appear in the definition of the vertex operators, however, the automorphism group is actually an extension of $\cdot 0$ by the abelian group $(\mathbf{Z}_2)^{24}$, denoted by $2^{24}(\cdot 0)$. This group has an analog in the theory of Lie algebras of type A, D, E. There one finds that the order two reflections in simple roots that generate the Weyl group W lift to automorphisms of the Lie algebra which are of order four. This yields an automorphism group of the Lie algebra that is an extension of W by $(\mathbf{Z}_2)^r$, where r is the rank of the corresponding root lattice [18].

The group $2^{24}(\cdot 0)$ is very similar in structure to the centralizer of an involution that appears in the monster. (An involution, i, is an element of order two; its centralizer in the monster consists of all elements g of the monster for which $g^{-1}ig=i$.) This centralizer, denoted by $C=2_+^{1+24}(\cdot 1)$, is given by an extension of $\cdot 1=\cdot 0/\mathbf{Z}_2$ (the simple Conway group) by an "extra-special" group denoted 2_+^{1+24}, the latter having a 2^{12}-dimensional irreducible representation. The group 2_+^{1+24} is isomorphic to the $\mathbf{Z}_2$ central extension of $\Lambda/2\Lambda$ discussed previously. The extra-special groups that arise in this way from $\mathbf{Z}_2$ twists have simple representations constructed in terms of γ-matrices (see e.g. [17] for an explicit construction). For example, a $\mathbf{Z}_2$ twist of the $SU(3)$ root lattice gives rise to an extra-special group that has a two-dimensional representation given by the conventional Pauli σ-matrices. These facts inspired FLM to consider a $\mathbf{Z}_2$ twisted version of the Leech theory in order to remove the unwanted 24 states that contribute to the constant term in the partition function, and to obtain the appropriate finite group structure. From an orbifold point of view, the centralizer of an involution

in the monster has a very suggestive form. The group $\cdot 1$ can be regarded as the automorphism group of the Leech lattice with lattice elements identified with their negatives (i.e. after identifying $x \to -x$), and, as seen in the last section, 2^{12} is the degeneracy factor required in the twisted sector for a purely left-acting $\mathbf{Z}_2$ twist of 24 coordinates. The involution i itself will turn out to be the symmetry of the operator algebra of a $\mathbf{Z}_2$ orbifold in which untwisted (twisted) states are assigned $+1$ (-1).

We now construct explicitly the $\mathbf{Z}_2$ twisted version of an arbitrary $c=24$ holomorphic conformal theory based on $\mathbf{R}^{24}/\Lambda$, where Λ could be any of the 24 even, self-dual, Niemeier lattices in 24 dimensions. The conformal field theory of interest for the monster will then be given by the choice $\Lambda = \Lambda_{\text{Leech}}$. With $g: x \to -x$, the untwisted Hilbert space $H_{(+)}$ decomposes into $g = \pm 1$ eigenspaces as

$$\begin{aligned} H^+_{(+)} &= \{\alpha^{i_1}_{-n_1} \cdots \alpha^{i_{2k}}_{-n_{2k}}(|\beta\rangle + |-\beta\rangle)\} + \{\alpha^{i_1}_{-n_1} \cdots \alpha^{i_{2k+1}}_{-n_{2k+1}}(|\beta\rangle - |-\beta\rangle)\}, \\ H^-_{(+)} &= \{\alpha^{i_1}_{-n_1} \cdots \alpha^{i_{2k+1}}_{-n_{2k+1}}(|\beta\rangle + |-\beta\rangle)\} + \{\alpha^{i_1}_{-n_1} \cdots \alpha^{i_{2k}}_{-n_{2k}}(|\beta\rangle - |-\beta\rangle)\} \end{aligned} \tag{5.1}$$

($n_i \in \mathbf{Z}, \beta \in \Lambda$), and we find

$$\begin{aligned} Z^+_{(+)}(q) &= q^{-1}\,\mathrm{Tr}_{H^+_{(+)}} q^{L_0} = \frac{1}{2}\left(\frac{\Theta_\Lambda(q)}{\eta^{24}} + \frac{q^{-1}}{\prod_{n=1}^{\infty}(1+q^n)^{24}}\right), \\ Z^-_{(+)}(q) &= q^{-1}\,\mathrm{Tr}_{H^-_{(+)}} q^{L_0} = \frac{1}{2}\left(\frac{\Theta_\Lambda(q)}{\eta^{24}} - \frac{q^{-1}}{\prod_{n=1}^{\infty}(1+q^n)^{24}}\right). \end{aligned} \tag{5.2}$$

The twisted Hilbert space $H_{(-)}$ decomposes as

$$\begin{aligned} H^+_{(-)} &= \{\alpha^{i_1}_{-n_1} \cdots \alpha^{i_{2k+1}}_{-n_{2k+1}}|0\rangle \otimes |\sigma^a\rangle\}, \\ H^-_{(-)} &= \{\alpha^{i_1}_{-n_1} \cdots \alpha^{i_{2k}}_{-n_{2k}}|0\rangle \otimes |\sigma^a\rangle\} \end{aligned} \tag{5.3}$$

($n_i \in \mathbf{Z} + \frac{1}{2}, |\sigma^a\rangle \in T$), so that

$$\begin{aligned} Z^+_{(-)}(q) &= q^{-1}\,\mathrm{Tr}_{H^+_{(-)}} q^{L_0} = q^{1/2} 2^{12} \frac{1}{2}\left(\frac{1}{\prod_{n=1}^{\infty}(1-q^{n-1/2})^{24}} - \frac{1}{\prod_{n=1}^{\infty}(1+q^{n-1/2})^{24}}\right), \\ Z^-_{(-)}(q) &= q^{-1}\,\mathrm{Tr}_{H^-_{(-)}} q^{L_0} = q^{1/2} 2^{12} \frac{1}{2}\left(\frac{1}{\prod_{n=1}^{\infty}(1-q^{n-1/2})^{24}} + \frac{1}{\prod_{n=1}^{\infty}(1+q^{n-1/2})^{24}}\right). \end{aligned} \tag{5.4}$$

The partition function for the twisted theory is thus given by

$$\begin{aligned} Z_{\text{tw}}(q) &= Z^+_{(+)}(q) + Z^+_{(-)}(q) = \frac{1}{2}\frac{\Theta_\Lambda(q)}{\eta^{24}} + \tfrac{1}{2} 2^{12}\left[\left(\frac{\eta}{\vartheta_2}\right)^{12} + \left(\frac{\eta}{\vartheta_4}\right)^{12} - \left(\frac{\eta}{\vartheta_3}\right)^{12}\right] \\ &= \frac{1}{2}\frac{\Theta_\Lambda(q)}{\eta^{24}} + \frac{1}{2}\frac{1}{\eta^{24}}[(\vartheta_3\vartheta_4)^{12} + (\vartheta_2\vartheta_3)^{12} - (\vartheta_2\vartheta_4)^{12}] \\ &= \tfrac{1}{2}(J(q) + 24(h+1)) + \tfrac{1}{2}(J(q) - 24) = J + 12h, \end{aligned} \tag{5.5}$$

where $h = |\Lambda_2|/24$ is the Coxeter number of the lattice, and $\vartheta_i = \vartheta_i(0,\tau)$ are the usual Jacobi theta-functions. The last line of (5.5) follows from the fact that $Z_{\rm tw}$ is a modular invariant function with a single pole with unit residue at $q=0$ (the cusp $\tau = i\infty$) and otherwise holomorphic, and it must therefore equal the modular function J up to an additive constant. For $\Lambda = \Lambda_{\rm Leech}$ as considered in (4.11), we have $\Lambda_2 = \emptyset$, and (5.5) gives

$$Z^{\rm tw}_{\rm Leech}(q) = Z_{\rm Leech}(q) - 24 = J(q). \tag{5.6}$$

We see that the twisting procedure preserves modular invariance while removing the 24 unwanted states $\alpha^i_{-1}|0\rangle$.

The twisted theory contains 196884 dimension 2 operators, including the stress tensor $T = -\frac{1}{2}\partial x\cdot\partial x$. In the untwisted Hilbert space there are $24\cdot 25/2$ operators $\partial x^i \partial x^j$ and $196560/2$ more dimension 2 operators $(e^{ip\cdot x} + e^{-ip\cdot x})/\sqrt{2}$ ($p \in \Lambda_4$). The remaining dimension 2 operators are the $24\cdot 2^{12}$ operators τ^{ia} (see (4.24)) that create states $\alpha^i_{-1/2}|\sigma^a\rangle$ in the twisted Hilbert space. FLM define a closed algebra in terms of the Fourier components ϕ_n, of these dimension 2 fields by defining a "cross-bracket,"

$$\phi^i_m \times \phi^j_n = \tfrac{1}{2}([\phi^i_{m+1}, \phi^j_{n-1}] + [\phi^j_{n+1}, \phi^i_{m-1}]), \tag{5.7}$$

where i, j label different dimension 2 fields. They then show that this cross bracket of dimension 2 fields closes on the Fourier components of other dimension 2 fields. This gives a commutative, non-associative, vertex operator algebra only slightly similar to the more familiar affine Lie algebras constructed from fields of dimension 1.

The above construction is easily understood from the OPE point of view. We have a conformal field theory with no dimension 1 fields, primary fields ϕ^i of dimension 2, a stress tensor $T(z)$ which is a descendant of the identity operator, and some set of primary dimension 3 fields χ^i. The OPE of $\phi^i(z)$ with $\phi^j(w)$ then has the general form

$$\begin{aligned}\phi^i(z)\phi^j(w) \sim{}& \frac{1}{(z-w)^4}\delta^{ij} + \frac{1}{(z-w)^2}\left(b^{ijk}\phi^k(w) + \frac{4}{c}\delta^{ij}T(w)\right)\\ &+ \frac{1}{(z-w)}\left(c^{ijk}\chi^k(w) + \tfrac{1}{2}\partial_w\left(b^{ijk}\phi^k(w) + \frac{4}{c}\delta^{ij}T(w)\right)\right) + \cdots,\end{aligned} \tag{5.8}$$

where Bose statistics requires symmetry of the coefficients b^{ijk} in the first two indices i, j, and antisymmetry of the c^{ijk} in i, j. The commutators in (5.7) can be extracted as usual from the singular terms in the OPE (5.8). In the cross bracket, the terms involving χ^i then cancel due to antisymmetry of the c^{ijk}.

Specializing to the case $c = 24$, we obtain (using also the OPE's (2.3) and (2.4))

$$\begin{aligned}&\phi^i_m \times \phi^j_n = b^{ijk}\phi^k_{m+n} + \tfrac{1}{6}\delta^{ij}L_{m+n} + \frac{m^2}{2}\delta^{ij}\delta_{m+n,0},\\ &L_m \times \phi^i_m = 2\phi^i_{m+n},\\ &L_m \times L_n = 2L_{m+n} + 6m^2\delta_{m+n,0}.\end{aligned} \tag{5.9}$$

FLM have shown that the infinite dimensional closed algebra defined by this "cross-bracket" gives an affinization of the Griess algebra B, which appears as the dimension 196884 zero mode subalgebra (i.e. represented by the ϕ_0^i's). The automorphism group of the Griess algebra is known to be precisely the monster [19]. From a "physical" point of view, the monster acts on this conformal field theory as a symmetry, commuting with the Hamiltonian and leaving the operator product coefficients invariant. The affinization of the Griess algebra is an example of a rather enlarged chiral algebra.

To give a better feeling for the monster, we recall that it can be generated by the centralizer C of the involution i together with another involution denoted σ [19]. As mentioned at the beginning of this section, the involution i is here the generic $\mathbf{Z}_2$ orbifold symmetry operation that acts as $+1$ on untwisted and -1 on twisted states. Because the group C commutes with i, each of its elements takes untwisted to untwisted states, and twisted to twisted states. The action of the additional symmetry σ, on the other hand, exchanges twisted and untwisted states. To identify the symmetry σ in our language, it is useful to regard the monster conformal field theory not as a $\mathbf{Z}_2$ orbifold of the conformal field theory based on the Leech lattice, but rather as a $\mathbf{Z}_2 \times \mathbf{Z}_2$ orbifold of the conformal field theory based on the Niemeier lattice $\Lambda_{A_1^{24}}$. (This is a lattice constructed by adjoining to the $SU(2)^{24}$ root lattice additional weight vectors specified by the Golay code (see e.g. [20]).) The two elements generating the $\mathbf{Z}_2 \times \mathbf{Z}_2$ group are $g: x \to -x$ and

$$h: x \to x + 2\pi \cdot \frac{1}{4} \sum_{i=1}^{24} \alpha_i, \tag{5.10}$$

where α_i are the 24 positive roots of $SU(2)^{24}$. The $\mathbf{Z}_2 \times \mathbf{Z}_2$ orbifold can be constructed in two steps, first modding out by the $\mathbf{Z}_2$ generated by the translation h, and then by the $\mathbf{Z}_2$ generating by g. The first step simply changes the lattice on which $x(z)$ is defined from $\Lambda_{A_1^{24}}$ to Λ_{Leech} [1], and the second by definition gives the monster conformal field theory.

Acting directly on the $\Lambda_{A_1^{24}}$ theory, on the other hand, g, h, and gh are all equivalent as $\mathbf{Z}_2$ twists. This may be seen by considering their actions on the level one $SU(2)^{24}$ current algebra of the theory,

$$\begin{aligned} g&: \ J_1^i \to +J_1^i, \quad J_2^i \to -J_2^i, \quad J_3^i \to -J_3^i, \\ h&: \ J_1^i \to -J_1^i, \quad J_2^i \to -J_2^i, \quad J_3^i \to +J_3^i, \\ gh&: \ J_1^i \to -J_1^i, \quad J_2^i \to +J_2^i, \quad J_3^i \to -J_3^i, \end{aligned} \tag{5.11}$$

$(i = 1, \ldots, 24)$ where

$$J_1^i = \frac{1}{\sqrt{2}}(e^{i\sqrt{2}x^i} + e^{-i\sqrt{2}x^i}), \quad J_2^i = \frac{i}{\sqrt{2}}(e^{i\sqrt{2}x^i} - e^{-i\sqrt{2}x^i}), \quad J_3^i = i\partial x^i \tag{5.12}$$

represent the currents. The three symmetry actions in (5.11) are equivalent due to the S_3 permutation symmetry of $SU(2)$ that interchanges the 3 currents. This triality symmetry is a symmetry of the full theory due to the $SU(2)^{24}$ symmetry. Because g, h and gh have equivalent $\mathbf{Z}_2$ actions, the $\mathbf{Z}_2 \times \mathbf{Z}_2$ orbifold conformal field theory also has an S_3 permutation symmetry that permutes fields in the 3 twisted sectors

H_g, H_h, and H_{gh} among one another, and also permutes fields within the untwisted sector H_1. The (odd) permutation that exchanges g and h is the symmetry σ that together with C generates the monster.

6. Superconformal Symmetry in the Moonshine Module

It is now straightforward to identify the superconformal symmetry in the FLM construction. The direct sum of the untwisted and twisted Hilbert spaces, $H_{(+)} \oplus H_{(-)}$, before g projection, gives rise to a $\mathbf{Z}_2$ non-local covering theory that contains 2^{12} primary dimension 3/2 twist fields σ^a. In general, the 2^{12} states $|\sigma^a\rangle$ (belonging to T, as discussed after (4.16)) are linear combinations of the 2^{24} states $|\tilde{q}_b\rangle$ located at fixed points. However, it is always possible [16] to take one of the states, say $|\sigma^0\rangle$, to be located at any particular fixed point, say the origin. The remaining linear combinations of states $|\tilde{q}_b\rangle$ are then generated by taking operator products of the distinguished twist field σ^0 with all the $\mathbf{Z}_2$-invariant combinations of exponentials (along with their cocycles) from the untwisted sector. We tentatively identify σ^0 (up to normalization) with the field T_F of a superconformal field theory. In order to check that (3.1) is satisfied, we need to determine the OPE of $\sigma^0(z)$ with $\sigma^0(w)$. Fields $\mathcal{O}(w)$ appearing in this OPE must create states in the untwisted Hilbert space and must have non-vanishing three-point function $\langle \sigma^0(z)\mathcal{O}(y)\sigma^0(w)\rangle \neq 0$. These three-point functions are determined by $SL(2,\mathbf{R})$ invariance, up to an overall factor that can be determined by calculating the expectation value $\langle \sigma^0|\mathcal{O}|\sigma^0\rangle$ of $\mathcal{O}(y)$ between twisted ground states $\langle \sigma^0| = \langle 0|\sigma^0(\infty)$ and $|\sigma^0\rangle = \sigma^0(0)|0\rangle$.

The space group selection rules [21] show that the only exponentials $e^{i\beta\cdot x(z)}$ appearing in the OPE of $\sigma^0(z)$ with $\sigma^0(w)$ have $\beta \in 2\Lambda$, and none of these has dimension 2 (required for a singular contribution to the OPE). The three-point functions needed to determine the coefficient of $\partial x^i \partial x^j$ in this OPE can be easily calculated using the expansion (4.16), with the result that this operator appears only for $i = j$, and with coefficient independent of j. Thus the only dimension 2 operator that has a non-vanishing expectation value between the same twisted ground state is the stress tensor $T = -\frac{1}{2}\sum_{i=1}^{24} \partial x^i \partial x^i$. The coefficient of T in the OPE is determined by associativity of the $T\sigma\sigma$ OPE (given that σ^0 is primary with dimension 3/2 and canonically normalized, as in (4.26)) to be $2\cdot(3/2)/c = 1/8$. We see that $T_F = 2\sigma_0$ satisfies (3.1) and can therefore serve as a supersymmetry generator.

We can now organize the primary fields into Ramond and Neveu–Schwarz sectors. Fields in the R sector have square-root type branch cuts in their OPE with T_F. From (4.24) we see that the Ramond fields are those that create states in $H^+_{(-)}$ or $H^-_{(+)}$. Similarly, NS fields create states in $H^+_{(+)}$ or $H^-_{(-)}$. The $\mathbf{Z}_2$ twist generator g plays the role that would be played by $\hat{\Gamma}$ in a general superconformal theory, in that projection onto the g invariant states $H^+_{(+)} \oplus H^+_{(-)}$ gives a local, modular invariant theory. The 24 states $\alpha^i_{-1}|0\rangle \in H^-_{(+)}$ created by the untwisted fields ∂x^i (which are Ramond fields with respect to $T_F = 2\sigma^0$) are Ramond ground states with $L_0 = \hat{c}/16 = 1$. G_0 pairs the $24\cdot 2^{12} = 98304$ ($\hat{\Gamma} = 1$) dimension 2 Ramond

fields τ^{ia} with the $196560/2 + 24 = 98304$ ($\hat{\Gamma} = -1$) fields $(e^{ip\cdot x} - e^{-ip\cdot x})/i\sqrt{2}$ and $\partial^2 x^i$. This pairing of Ramond states into pairs with opposite eigenvalue of $\hat{\Gamma} = g$ is easily verified at all higher levels. To see this note that $q^{-1}\,\mathrm{Tr}_{H^-_{(+)}} q^{L_0} - 24 = q^{-1}\,\mathrm{Tr}_{H^+_{(-)}} q^{L_0}$, as follows from the identities implicit in (5.5) (in general we have $Z^-_{(+)}(q) - Z^+_{(-)}(q) = 12h + 24$).

In the NS sector, on the other hand, superconformal invariance organizes the fields into primary superfields. The identity operator is a NS field of dimension 0. It gives rise to a dimension 3/2 descendant $G_{-3/2}|0\rangle$ and a dimension 2 descendant $L_{-2}|0\rangle$ (recall that the vacuum is annihilated by $G_{-1/2}$ and L_{-1}). The remaining $2^{12} - 1$ of the dimension 3/2 NS fields σ^a are lowest components of primary superfields of dimension 3/2. The highest components of these superfields are the descendants with respect to $G_{-1/2}$ of the lowest components, i.e. they correspond to the states $G_{-1/2}|\sigma^a\rangle$. Subtracting these out, we are left with $196560/2 + 300 - 2^{12} = 94484$ primary superfields of dimension 2.

In general, the NS primary superfields can be classified according to their dimension and the $\hat{\Gamma}$ eigenvalue of their lowest component. To count primary superfields we must subtract all the states that are descendants under the superconformal algebra. Since the identity operator is annihilated by $G_{-1/2}$ and L_{-1}, it must be treated specially. We first subtract from $Z^+_{(+)}$ and $Z^-_{(-)}$ the descendants of the identity with $\hat{\Gamma}$ eigenvalue $+1$ and -1 respectively. This yields

$$Y^+_{(+)}(q) = Z^+_{(+)}(q) - \frac{\frac{1}{2}\left(\prod_{n=2}^{\infty}(1+q^{n-1/2}) + \prod_{n=2}^{\infty}(1-q^{n-1/2})\right)}{\prod_{n=2}^{\infty}(1-q^n)}, \tag{6.1}$$

$$Y^-_{(-)}(q) = Z^-_{(-)}(q) - \frac{\frac{1}{2}\left(\prod_{n=2}^{\infty}(1+q^{n-1/2}) - \prod_{n=2}^{\infty}(1-q^{n-1/2})\right)}{\prod_{n=2}^{\infty}(1-q^n)}. \tag{6.2}$$

We let

$$P_{\pm}(q) = q^{-1}\sum_{h>0} N_{\pm}(h) q^h \tag{6.3}$$

be the generating function for primary NS superfields (other than the identity) with $\hat{\Gamma}$ eigenvalue ± 1, where $N_{\pm}(h)$ is the number of NS primary superfields of dimension h and $\hat{\Gamma}$ eigenvalue ± 1. Writing $Y^+_{(+)}$ and $Y^-_{(-)}$ in terms of primary fields and their descendants and then solving for $P_{\pm}$ gives

$$P_{\pm}(q) = \frac{1}{2}\prod_{n=1}^{\infty}(1-q^n)\left[\frac{Y^+_{(+)}(q) + Y^-_{(-)}(q)}{\prod_{n=1}^{\infty}(1+q^{n-1/2})} \pm \frac{Y^+_{(+)}(q) - Y^-_{(-)}(q)}{\prod_{n=1}^{\infty}(1-q^{n-1/2})}\right]. \tag{6.4}$$

While it is relatively easy to count the number of primary superfields at a given level, it is more difficult to determine precisely which combinations of fields correspond to these superfields. This involves a more detailed analysis of the structure of the Leech lattice and the two-cocycles $\varepsilon(\alpha, \beta)$ than will be attempted here. In

any event we see how the numerology associated with the coefficients of the expansion of the modular invariant J of (4.11) may be further refined by recognition of the superconformal symmetry in the FLM construction.

Now that we have identified the superconformal symmetry in the FLM construction, it is possible to give a superconformal interpretation of the removal of the 24 unwanted states from the Leech theory (4.11). Every superconformal theory has a $\mathbf{Z}_2$ symmetry that we denote $(-1)^{F_s}$, defined to act as $+1$ on states in the antiperiodic (NS) sector of the worldsheet supersymmetry generator, and as -1 on states in the periodic (R) sector. (We refer to this symmetry as $(-1)^{F_s}$ since it resembles the action of an operator $(-1)^{F_s}$ in string theory, where F_s would be the spacetime fermion number.) In Appendix A of [10], it was shown that modding out by this symmetry (i.e. take $g=(-1)^{F_s}$ in (4.23)) gives another superconformal theory with opposite sign for the term (3.5). (Modding out the resulting theory again by $(-1)^{F_s}$ gives back the original theory.) Modding out the twisted Leech theory by $(-1)^{F_s}$ thus gives back a superconformal theory in which the 24 Ramond ground states are restored. This theory is in fact the original Leech theory Z_{Leech}, which we now recognize to possess a superconformal symmetry as well. One way to see this is to recognize that the action of $(-1)^{F_s}$ on the states of the twisted Leech theory ($+1$ on $H^+_{(+)}$ and -1 on $H^+_{(-)}$) is identical to that of the $\mathbf{Z}_2$ symmetry that would generically undo the orbifolding by $x \to -x$ (i.e. the involution i, defined to act as $+1$ on the untwisted sector, and -1 on the twisted sector). In superconformal language, then, the modification that removes the 24 unwanted states from the Leech theory can be equally well taken to be the $\mathbf{Z}_2$ symmetry generated by $(-1)^{F_s}$, and the argument preceding (3.5) gives the natural explanation in this language for why the partition function is changed by a constant.

Having introduced the symmetry $(-1)^{F_s}$, we can now complete the discussion at the end of Sect. 3 concerning when a given conformal theory is superconformal. A key role in this identification is played by the $\mathbf{Z}_2$ symmetry $(-1)^{F_s}$, because it is a symmetry of *any* superconformal field theory (holomorphic or not). If a modular invariant superconformal theory is twisted by this $\mathbf{Z}_2$ symmetry, the operator T_F appears in the twisted Hilbert space $H_{(-)}$ (i.e. before projection onto $\mathbf{Z}_2$-invariant states—the final $\mathbf{Z}_2$ projection, giving a modular invariant partition function, would again eliminate T_F). To determine whether a conformal field theory is superconformal or not, one may thus twist by all $\mathbf{Z}_2$ symmetries of the operator algebra, and in each case look for a dimension $(\frac{3}{2}, 0)$ field in the twisted sector that obeys (3.1), and also has the appropriate local behavior with respect to the $\mathbf{Z}_2$-even (Neveu–Schwarz) fields and a square-root type non-local behavior with respect to the $\mathbf{Z}_2$-odd (Ramond) fields. If there is such a symmetry, then the combined untwisted plus twisted Hilbert space before $\mathbf{Z}_2$ projection, $H_{(+)} \oplus H_{(-)}$, provides the desired non-local $\mathbf{Z}_2$ covering theory. The $\mathbf{Z}_2$ symmetry that projects back onto the original modular invariant theory, now recognized as superconformal, is simply that assigning $+1$ (-1) to states in $H_{(+)}(H_{(-)})$.

It is easy to see that the supersymmetry in the twisted Leech theory terminates at $N=1$. This is because higher supersymmetry algebras involve as well dimension one currents not present in the theory (5.6). The monster group, although an automorphism group of the full conformal field theory, does not preserve the

supersymmetry generator, and the supersymmetric structure we have identified is invariant only under a subgroup (contained in the subgroup C) of the monster group.

7. The Monster and Genus Zero Function Fields

Perhaps the most mysterious element in the monstrous game is the relation between elements of the monster and genus zero function fields. In this section we will give a brief summary (again oriented towards string theorists) of what is known about this relation. A more detailed discussion can be found in [2, 22].

The fundamental domain for the modular group, $\mathbf{H}/SL(2,\mathbf{Z})$ (where $\mathbf{H}$ is the upper half-plane) is a familiar object to string theorists. After adjoining the point at infinity ($\tau = i\infty$), this space can be given a topology and a complex structure so that it becomes a compact Riemann surface. The modular function j provides a map of this compactification of $\mathbf{H}/SL(2,\mathbf{Z})$ to the Riemann sphere $\mathbf{C}\cup\infty$ which is one-to-one and onto. As a result, we see that the Riemann surface associated with $\mathbf{H}/SL(2,\mathbf{Z})$ has genus zero. It follows that the field $K_{SL(2,\mathbf{Z})}$ of meromorphic functions on this Riemann surface consists of all rational functions of j with complex coefficients, i.e. any $k\in K_{SL(2,\mathbf{Z})}$ satisfies

$$k = \frac{\sum_{i=0}^{n} a_i j^i}{\sum_{i=0}^{m} b_i j^i} \tag{7.1}$$

with $a_i, b_i \in \mathbf{C}$. $K_{SL(2,\mathbf{Z})}$ is said to be a genus zero function field, and the modular function j is the generator, or *hauptmodul* of this function field.

This discussion can be repeated for other discrete subgroups $\Gamma \subset SL(2,\mathbf{R})$, known as groups of genus zero, that have the property that the appropriate compactification of $\mathbf{H}/\Gamma$ yields a Riemann surface of genus zero. The field of rational functions on the compactification of $\mathbf{H}/\Gamma$ is a genus zero function field with generator j_Γ, analogous to the modular function j in the case $\Gamma = SL(2,\mathbf{Z})$. Examples of such genus zero subgroups of $SL(2,\mathbf{R})$ are provided by the congruence subgroups $\Gamma_0(p)$, where p is a prime such that $p-1$ divides 24. Here $\Gamma_0(N)$ is defined as

$$\Gamma_0(N) = \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in SL(2,\mathbf{Z}) \;\middle|\; c = 0 \bmod N \right\}. \tag{7.2}$$

Another group of interest, denoted $\Gamma_0(N)^+$, is obtained by adjoining the Fricke involution $\tau \to -1/N\tau$ to $\Gamma_0(N)$.[2] It was observed in [24] that the primes p for which $\Gamma_0(p)^+$ has genus zero are precisely the primes that divide the order of the monster!

Since the monster acts as a symmetry of the (super) conformal field theory constructed by FLM, it has a well-defined action on each graded subspace, and

[2] See [23] for an application of these groups in a string theory context

one can define the series

$$T_g(q) = q^{-1}\,\mathrm{Tr}\, g q^{L_0}, \tag{7.3}$$

where g is an element of the monster, and the trace runs over all states in the theory. These series are known as Thompson–McKay series. Conway and Norton have conjectured that the Thompson–McKay series associated to each rational conjugacy class of the monster is a hauptmodul for a genus zero subgroup $\Gamma \subset SL(2, \mathbf{R})$ that contains $\Gamma_0(N)$ and is contained in the normalizer of $\Gamma_0(N)$ in $SL(2, \mathbf{R})$. (The normalizer of a subgroup $H \subset G$ consists of all elements $g \in G$ such that $g^{-1}hg \in H$ for all $h \in H$. A rational conjugacy class is one whose character is a rational number.)

The Thompson series $T_g(q)$ can be viewed as a path integral over the torus with modular parameter τ and boundary conditions given by $g\underset{1}{\square}$. In the spirit of orbifolds, it is natural to consider more general boundary conditions on the torus. If g and h are two commuting elements of the monster, then it makes sense to consider the path integral (4.21) with boundary conditions twisted by g and h in the "time" and "space" directions. This corresponds to the trace

$$Z(g, h; q) \equiv g\underset{h}{\boxed{}} = q^{-1}\,\mathrm{Tr}_{H_h}\, g q^{L_0}, \tag{7.4}$$

evaluated in the Hilbert space twisted by h. The standard orbifold result that $Z(g, h; q)$ is invariant under simultaneous conjugation of both g and h, and the result that the boundary conditions (4.21) transform as in (4.22) under modular transformations, appear as parts one and three of the conjecture by Norton in the appendix of [22]. Norton also conjectures that these generalized Thompson–McKay series are invariant under genus zero subgroups of $SL(2, \mathbf{R})$. This conjecture strengthens our suspicion that holomorphic orbifold (super) conformal field theories may provide the proper setting for trying to obtain a conceptual understanding of the relations between the monster and genus zero function fields.

8. Discussion

It would be interesting to investigate other orbifold holomorphic conformal field theories in an attempt to generalize the FLM construction of the monster to other sporadic groups, and to study the structure of twisted vertex operator algebras. It is easy using (5.5) to check that $\mathbf{Z}_2$ twists of the other 23 Niemeier lattices lead in some cases to holomorphic conformal field theories which are not graded isomorphic to any of the untwisted theories based on Niemeier lattices. It might be illuminating to investigate the structure of the cross-bracket algebra in these cases as well. We point out that the arguments given in the second paragraph of Sect. 6, showing that the twist field is a supersymmetry generator, apply equally well to $\mathbf{Z}_2$ twisted versions of the $c_1 = 24$ holomorphic conformal field theories based on the remaining 23 Niemeier lattices. According to (5.5), the $x \to -x$ modification always removes $12h + 24$ states, which may then be restored by modding out the resulting superconformal theory by $(-1)^{F_s}$, thereby manifesting

as well a superconformal invariance in the untwisted Niemeier theories. Since the Niemeier conformal theories in general possess spin one states in their spectra, we suspect the occurrence of higher ($N > 1$) superconformal algebras in many cases (depending on the detailed structure of the cocycles).

It is also possible that one may need to consider holomorphic conformal field theories invariant only under a subgroup of the modular group. In [1], for example, FLM provide some evidence that a twisted version of compactification of the left-moving superstring on the E_8 lattice also has moonshine properties for $\cdot 1$. This theory also has an underlying superconformal symmetry, but again it is not clear what role, if any, it plays in the moonshine.

The main outstanding problems related to the moonshine module are the understanding of the relation between elements of the monster and genus-zero function fields; and a generalization of the techniques reviewed here to other sporadic groups, particularly the "pariahs" which are not involved in the monster. We reiterate our hope that the techniques of (super) conformal field theory will play a useful role in the solution to these problems.

Acknowledgements. L. D. is supported by NSF contract PHY80-19754; P. G. by NSF contract PHY-82-15249, DOE OJI grant FG-84ER40171, and by the A. P. Sloan foundation; J. H. by NSF contract PHY80-19754, PYI award PHY86-58033, and by the A. P. Sloan foundation. We are grateful to I. Frenkel, J. Lepowsky, and G. Moore for comments on the manuscript, and to V. Kaplunovsky for a useful discussion.

References

1. Frenkel, I. B., Lepowsky, J., Meurman, A.: A natural representation of the Fischer-Griess monster with the modular function J as character. Proc. Natl. Acad. Sci. U.S.A. **81**, 32566 (1984); Frenkel, I. B., Lepowsky, J., Meurman, A.: A moonshine module for the monster. In: Vertex operators in mathematics and physics. Lepowsky, J., Mandelstam, S., Singer, I (eds.). Publ. Math. Sciences Res. Inst. Vol. 3, p. 231. Berlin, Heidelberg, New York: Springer 1985; Frenkel, I. B., Lepowsky, J., Meurman, A.: Vertex operator algebras and the monster. New York: Academic Press (to appear)
2. Conway, J. H., Norton, S. P.: Monstrous moonshine. Bull. Lond. Math. Soc. **11**, 308 (1979)
3. Belavin, A. A., Polyakov, A. M., Zamolodchikov, A. B.: Infinite conformal symmetry in two-dimensional quantum field theory. Nucl. Phys. **B241**, 333 (1984); Friedan, D.: Notes on string theory and two dimensional conformal field theory. In: Unified string theories. Green, M., Gross, D. (eds.). Singapore: World Scientific 1986; Peskin, M.: Introduction to string and superstring theory II. 1986 TASI lectures, SLAC-PUB-4251; Banks, T.: Lectures on conformal field theory. 1987 TASI lectures, SCIPP 87/111
4. Friedan, D., Qiu, Z., Shenker, S.: Conformal invariance, unitarity, and critical exponents in two dimensions. Phys. Rev. Lett. **52**, 1575 (1984)
5. Cardy, J.: Operator content of two-dimensional conformally invariant theories. Nucl. Phys. **B270**[FS16], 186 (1986)
6. Moore, G.: private communication
7. Gliozzi, F., Scherk, J., Olive, D.: Supersymmetry, supergravity theories and the dual spinor model. Nucl. Phys. **B122**, 253 (1977); Friedan, D., Qiu, Z., Shenker, S.: Superconformal invariance in two dimensions and the tricritical Ising model. Phys. Lett. **151B**, 37 (1985)
8. Witten, E.: Constraints on supersymmetry breaking. Nucl. Phys. **B202**, 253 (1982)
9. Kastor, D.: Modular invariance in superconformal models. Nucl. Phys. **B280**[FS18] 304 (1987)
10. Dixon, L., Ginsparg, P., Harvey, J.: $\hat{c} = 1$ superconformal field theory. Nucl. Phys. **B306**, 470 (1988)

11. Narain, K. S., Sarmadi, M. H., Witten, E.: A note on toroidal compactification of heterotic string theory. Nucl. Phys. **B279** 369 (1986); Ginsparg, P.: On toroidal compactification of heterotic superstrings. Phys. Rev. **D35**, 648 (1987)
12. Frenkel, I. B., Kac, V. G.: Basic representations of affine Lie algebras and dual resonance models. Invent. Math. **62**, 23 (1980); Segal, G.: Unitary representations of some infinite dimensional groups. Commun. Math. Phys. **81**, 301 (1981)
13. Dixon, L., Harvey, J., Vafa, C., Witten, E.: Strings on orbifolds I, II. Nucl. Phys. **B261**, 678 (1985); Nucl. Phys. **B274** 285 (1986)
14. Vafa, C.: Modular invariance and discrete torsion on orbifolds. Nucl. Phys. **B273**, 592 (1986)
15. Lepowsky, J.: Calculus of twisted vertex operators. Proc. Natl. Acad. Sci. USA **82**, 8295 (1985)
16. Narain, K. S., Sarmadi, M. H., Vafa, C.: Asymmetric orbifolds. Nucl. Phys. **B288**, 551 (1987)
17. Corrigan, E., Hollowood, T.: Comments on the algebra of straight, twisted, and intertwining vertex operators. Nucl. Phys. **B304**, 77 (1988)
18. Tits, J.: Normalisateurs de tores 1. Groups de Coxeter étendus. J. Alg. **4**, 96 (1966)
19. Griess, R. L., Jr.,: The friendly giant. Invent. Math. **69**, 1 (1982); Tits, J.: On R. Griess' 'Friendly Giant'. Invent. Math. **78**, 491 (1984)
20. Conway, J. H., Sloane, N. J. A.: Twenty-three constructions for the Leech lattice. Proc. Roy. Soc. Lond. Ser. **A381**, 275 (1982)
21. Dixon, L., Friedan, D., Martinec, E., Shenker, S.: The conformal field theory of orbifolds. Nucl. Phys. **B282**, 13 (1987); Hamidi, S., Vafa, C.: Interactions on orbifolds. Nucl. Phys. **B279**, 465 (1987)
22. Mason G. (with an appendix by Norton, S. P.),: Finite groups and modular functions. Proc. Symposia in Pure Math. **47**, 181 (1987)
23. Moore, G.: Atkin-Lehner symmetry. Nucl. Phys. **B293,** 139 (1987). Erratum ibid. **B299,** 847 (1988)
24. Ogg, A.: Hyperelliptic modular curves. Bull. Soc. Math. France **102**, 449 (1974)

Communicated by L. Alvarez-Gaumé

Received April 27, 1988

Monstrous moonshine and monstrous Lie superalgebras.

Invent. Math. 109, 405-444 (1992).

Richard E. Borcherds,

Department of pure mathematics and mathematical statistics, 16 Mill Lane, Cambridge CB2 1SB, England.

We prove Conway and Norton's moonshine conjectures for the infinite dimensional representation of the monster simple group constructed by Frenkel, Lepowsky and Meurman. To do this we use the no-ghost theorem from string theory to construct a family of generalized Kac-Moody superalgebras of rank 2, which are closely related to the monster and several of the other sporadic simple groups. The denominator formulas of these superalgebras imply relations between the Thompson functions of elements of the monster (i.e. the traces of elements of the monster on Frenkel, Lepowsky, and Meurman's representation), which are the replication formulas conjectured by Conway and Norton. These replication formulas are strong enough to verify that the Thompson functions have most of the "moonshine" properties conjectured by Conway and Norton, and in particular they are modular functions of genus 0. We also construct a second family of Kac-Moody superalgebras related to elements of Conway's sporadic simple group Co_1. These superalgebras have even rank between 2 and 26; for example two of the Lie algebras we get have ranks 26 and 18, and one of the superalgebras has rank 10. The denominator formulas of these algebras give some new infinite product identities, in the same way that the denominator formulas of the affine Kac-Moody algebras give the Macdonald identities.

1 Introduction.
2 Introduction (continued).
3 Vertex algebras.
4 Generalized Kac-Moody algebras.
5 The no-ghost theorem.
6 Construction of the monster Lie algebra.
7 The simple roots of the monster Lie algebra.
8 The twisted denominator formula.
9 The moonshine conjectures.
10 The monstrous Lie superalgebras.
11 Some modular forms.
12 The fake monster Lie algebra.
13 The denominator formula for fake monster Lie algebras.
14 Examples of fake monster Lie algebras.
15 Open problems.

1 Introduction.

The main result of the first half of this paper is the following.

Theorem 1.1. *Suppose that $V = \oplus_{n\in\mathbf{Z}} V_n$ is the infinite dimensional graded representation of the monster simple group constructed by Frenkel, Lepowsky, and Meurman [16,17]. Then for any element g of the monster the Thompson series $T_g(q) = \sum_{n\in\mathbf{Z}} \mathrm{Tr}(g|V_n)q^n$ is*

a Hauptmodul for a genus 0 subgroup of $SL_2(\mathbf{R})$, i.e., V satisfies the main conjecture in Conway and Norton's paper [13].

We prove this by constructing a $\mathbf{Z}^2$-graded Lie algebra acted on by the monster, called the monster Lie algebra. This is a generalized Kac-Moody algebra, and by calculating the "twisted denominator formulas" of this Lie algebra explicitly we get enough information about the Thompson series $T_g(q)$ to determine them.

In this introduction we explain this result in more detail and briefly describe the proof, which is contained in sections 6 to 9. The rest of this paper is organized as follows. Section 2 is an introduction to the second half of the paper (sections 10 to 14) which uses some of the techniques of the proof to find some new infinite product formulas and infinite dimensional Lie algebras. Sections 3, 4, and 5 summarize some known results about vertex algebras, Kac-Moody algebras, and the no-ghost theorem that we use in the proof of theorem 1.1. Section 15 contains a list of some open questions. There is a list of notation we use at the end of section 1.

The Fischer-Griess monster sporadic simple group, of order $2^{46}3^{20}5^9 7^6 11^2 13^3 17.19.23.29.31.41.47.59.71$, acts naturally and explicitly on a graded real vector space $V = \oplus_{n\in\mathbf{Z}} V_n$ constructed by Frenkel, Lepowsky and Meurman [16,17] (The vector spaces V and V_n are denoted by $V^\natural$ and $V^\natural_{-n}$ in [16].) The dimension of V_n is equal to the coefficient $c(n)$ of the elliptic modular function $j(\tau) - 744 = \sum_n c(n)q^n = q^{-1} + 196884q + 21493760q^2 + \dots$ (where we write q for $e^{2\pi i\tau}$, and $\mathrm{Im}(\tau) > 0$). One of the main remaining problems from [16], which theorem 1.1 solves, is to calculate the character of V as a graded representation of the monster, or in other words to calculate the trace $\mathrm{Tr}(g|V_n)$ of each element g of the monster on each space V_n. The best way to describe this information is to define the Thompson series

$$T_g(q) = \sum_{n\in\mathbf{Z}} \mathrm{Tr}(g|V_n)q^n$$

for each element g of the monster, so we want to calculate these Thompson series. For example, if 1 is the identity element of the monster then $\mathrm{Tr}(1|V_n) = \dim(V_n) = c(n)$, so that the Thompson series $T_1(q) = j(\tau) - 744$ is the elliptic modular function. McKay, Thompson, Conway and Norton conjectured [13] that the Thompson series $T_g(q)$ are all Hauptmoduls for certain explicitly given modular groups of genus 0. (More precisely, they only conjectured that there should be some graded module for the monster whose Thompson series are Hauptmoduls, since their conjectures came before the construction of V.) This conjecture follows from theorem 1.1. The corresponding Hauptmoduls are the ones listed in [13] (with their constant terms removed), so this completely describes V as a representation of the monster.

We recall the definition of a Hauptmodul. The group $SL_2(\mathbf{Z})$ acts on the upper half plane $H = \{\tau \in \mathbf{C} | \mathrm{Im}(\tau) > 0\}$ by $\left(\begin{smallmatrix} a & b \\ c & d\end{smallmatrix}\right)(\tau) = \frac{a\tau+b}{c\tau+d}$. A meromorphic function on H invariant under $SL_2(\mathbf{Z})$ and satisfying a certain regularity condition at $i\infty$ is called a modular function of level 1. The phrase "level 1" refers to the group $SL_2(\mathbf{Z})$; if this is replaced by some commensurable group we get modular functions of higher levels. The elliptic modular function $j(\tau)$ is, up to normalizations, the simplest nonconstant modular function of level 1; more precisely, the modular functions of level 1 are the rational functions

of j. The element $\left(\begin{smallmatrix}1 & 1\\ 0 & 1\end{smallmatrix}\right)$ of $SL_2(\mathbf{Z})$ takes τ to $\tau+1$, so in particular $j(\tau)$ is periodic and can be written as a Laurent series in $q=e^{2\pi i\tau}$. An exact expression for j is

$$j(\tau)=\frac{(1+240\sum_{n>0}\sigma_3(n)q^n)^3}{q\prod_{n>0}(1-q^n)^{24}}$$

where $\sigma_3(n)=\sum_{d|n}d^3$ is the sum of the cubes of the divisors of n; see any book on modular forms or elliptic functions, for example [30]. Another way of thinking about j is that it is an isomorphism from the quotient space $H/SL_2(\mathbf{Z})$ to the complex plane, which can be thought of as the Riemann sphere minus the point at infinity.

We can also consider functions invariant under some group G commensurable with $SL_2(\mathbf{Z})$ acting on H. The quotient H/G is again a compact Riemann surface $\overline{H/G}$ with a finite number of points removed. If this compact Riemann surface is a sphere, rather than something of higher genus, then we say that G is a genus 0 group. In this case a function giving an isomorphism from the compact Riemann surface $\overline{H/G}$ to the sphere $\mathbf{C}\cup\infty$ taking $i\infty$ to ∞ is called a Hauptmodul for the genus 0 group G; it is unique up to addition of a constant and multiplication by a nonzero constant. If a Hauptmodul can be written as $e^{-2\pi ia\tau}+$ a function vanishing at $i\infty$ for some positive a then we say that it is normalized. Every genus 0 group has a unique normalized Hauptmodul. For example, $j(\tau)-744=e^{-2\pi i\tau}+0+196884e^{2\pi i\tau}+\dots$ is the normalized Hauptmodul for the genus 0 group $SL_2(\mathbf{Z})$.

Another example is $G=\Gamma_0(2)$, where $\Gamma_0(N)=\{\left(\begin{smallmatrix}a & b\\ c & d\end{smallmatrix}\right)\in SL_2(\mathbf{Z})|c\equiv 0 \bmod N\}$. The quotient H/G is then a sphere with 2 points removed, so that G is a genus 0 group. Its normalized Hauptmodul is $T_{2-}(q)=24+q^{-1}\prod_{n>0}(1-q^{2n+1})^{24}=q^{-1}+276q-2048q^2+\dots$, and is equal to the Thompson series of a certain element of the monster of order 2 (of type 2B in atlas [14] notation). Similarly $\Gamma_0(N)$ is a genus 0 subgroup for several other values of N which correspond to elements of the monster. (However the genus of $\Gamma_0(N)$ tends to infinity as N increases, so there are only a finite number of integers N for which it has genus 0; more generally Thompson [31] has shown that there are only a finite number of conjugacy classes of genus 0 subgroups of $SL_2(\mathbf{R})$ which are commensurable with $SL_2(\mathbf{Z})$.)

So we want to calculate the Thompson series $T_g(\tau)$ and show that they are Hauptmoduls of genus 0 subgroups of $SL_2(\mathbf{R})$. The difficulty with doing this is as follows. Frenkel, Lepowsky, and Meurman constructed V as the sum of two subspaces V^+ and V^-, which are the $+1$ and -1 eigenspaces of a certain element of order 2 in the monster. If an element g of the monster commutes with this element of order 2, then it is not difficult to to work out its Thompson series $T_g(q)=\sum_n \mathrm{Tr}(g|V_n)q^n$ as the sum of two series given by its traces on V^+ and V^- (this is done in [16,17]), and it would probably be tedious but straightforward to check directly that these are all Hauptmoduls. Unfortunately, if an element of the monster is not conjugate to something that commutes with this element of order 2 then there is no obvious direct way of working out its Thompson series, because it muddles up V^+ and V^- in a very complicated way.

We calculate these Thompson series indirectly using the monster Lie algebra M. This is a $\mathbf{Z}^2=\mathbf{Z}\oplus\mathbf{Z}$ graded Lie algebra, whose piece of degree $(m,n)\in\mathbf{Z}^2$ is isomorphic as a module over the monster to V_{mn} if $(m,n)\neq(0,0)$ and to $\mathbf{R}^2$ if $(m,n)=(0,0)$, so for small

degrees it looks like

$$\begin{array}{cccccccccc}
 & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \\
\cdots & 0 & 0 & 0 & 0 & V_3 & V_6 & V_9 & \cdots \\
\cdots & 0 & 0 & 0 & 0 & V_2 & V_4 & V_6 & \cdots \\
\cdots & 0 & 0 & V_{-1} & 0 & V_1 & V_2 & V_3 & \cdots \\
\cdots & 0 & 0 & 0 & \mathbf{R}^2 & 0 & 0 & 0 & \cdots. \\
\cdots & V_3 & V_2 & V_1 & 0 & V_{-1} & 0 & 0 & \cdots \\
\cdots & V_6 & V_4 & V_2 & 0 & 0 & 0 & 0 & \cdots \\
\cdots & V_9 & V_6 & V_3 & 0 & 0 & 0 & 0 & \cdots \\
 & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots &
\end{array}$$

Very briefly, this Lie algebra is constructed as the space of physical states of a bosonic string moving in a Z_2-orbifold of a 26-dimensional torus (or strictly speaking, about half the physical states). See section 6 for more details. The space of physical states is a subquotient of a vertex algebra constructed from the vertex algebra V; vertex algebras are described in more detail in [3,16,18], and the properties we use are summarized in section 3. This subquotient can be identified using the no-ghost theorem from string theory ([21] or section 5), and is as described above.

We need to know what the structure of the monster Lie algebra is. It turns out to be something called a generalized Kac-Moody algebra, so we explain what these are.

Section 4 describes the results about generalized Kac-Moody algebras that we use. This paragraph gives a brief summary of them. Kac-Moody algebras can be thought of as Lie algebras generated by a copy of sl_2 for each point in their Dynkin diagram. Generalized Kac-Moody algebras are rather like Kac-Moody algebras except that we are allowed to glue together the sl_2's in more complicated ways, and are also allowed to use Heisenberg Lie algebras as well as sl_2's to generate the algebra. The main difference between Kac-Moody algebras and generalized Kac-Moody algebras is that the roots α of a Kac-Moody algebra may be either real ($(\alpha,\alpha) > 0$) or imaginary ($(\alpha,\alpha) \le 0$) but all the simple roots must be real, while generalized Kac-Moody algebras may also have imaginary simple roots. Kac-Moody algebras have a "denominator formula", which says that a product over positive roots is equal to a sum over the Weyl group; for example, the denominator formula for the affine Kac-Moody algebra $sl_2(\mathbf{R}[z,z^{-1}])$ is the Jacobi triple product identity. Generalized Kac-Moody algebras have a denominator formula which is similar to the one for Kac-Moody algebras, except that it has some extra correction terms for the imaginary simple roots. (The simple roots of a generalized Kac-Moody algebra correspond to a minimal set of generators for the subalgebra corresponding to the positive roots. For Kac-Moody algebras the simple roots also correspond to the points of the Dynkin diagram and to the generators of the Weyl group.)

We return to the monster Lie algebra. This is a generalized Kac-Moody algebra, and we will now write down its denominator formula, which says that a product over the positive roots is a sum over the Weyl group. The positive roots are the vectors (m,n) with $m > 0$, $n > 0$, and the vector $(1,-1)$, and the root (m,n) has multiplicity $c(mn)$. The Weyl group has order 2 and its nontrivial element maps (m,n) to (n,m), so it exchanges

$p = e^{(1,0)}$ and $q = e^{(0,1)}$. The denominator formula for the monster Lie algebra is the product formula for the j function

$$p^{-1} \prod_{m>0, n\in Z} (1 - p^m q^n)^{c(mn)} = j(p) - j(q).$$

(The left side is antisymmetric in p and q because of the factor of $p^{-1}(1 - p^1 q^{-1})$ in the product.) The reason why we get $j(p)$ and $j(q)$ rather than monomials in p and q on the right hand side (as we would for ordinary Kac-Moody algebras) is because of the correction due to the imaginary simple roots of M. The simple roots of M correspond to a set of generators of the subalgebra E of the elements of M whose degree is to the right of the y axis (so the roots of E are the positive roots of M), and turn out to be the vectors $(1, n)$ each with multiplicity $c(n)$. In the picture of the monster Lie algebra given earlier, the simple roots are given by the column just to the right of the one containing $\mathbf{R}^2$. The sum of the simple root spaces is isomorphic to the space V. The simple root $(1, -1)$ is real of norm 2, and the simple roots $(1, n)$ for $n > 0$ are imaginary of norm $-2n$ and have multiplicity $c(n) = \dim(V_n)$. As these multiplicities are exactly the coefficients of the j function, it is not surprising that j appears in the correction caused by the imaginary simple roots. This discussion is slightly misleading because we have implied that we obtain the product formula of the j function as the denominator formula of the monster Lie algebra by using our knowledge of the simple roots; in fact we really have to use this argument in reverse, using the product formula for the j function in order to work out what the simple roots of the monster Lie algebra are. We do this in section 7.

We can now extract information about the coefficients of the Thompson series $T_g(\tau)$ from a twisted denominator formula for the monster Lie algebra as follows: for an arbitrary generalized Kac-Moody algebra there is a more general version of the Weyl denominator formula which states that

$$\Lambda(E) = H(E),$$

where E is the subalgebra corresponding to the positive roots. Here $\Lambda(E) = \Lambda^0(E) \ominus \Lambda^1(E) \oplus \Lambda^2(E) \dots$ is a virtual vector space which is the alternating sum of the exterior powers of E, and similarly $H(E)$ is the alternating sum of the homology groups $H_i(E)$ of the Lie algebra E (see [10]). This identity is true for any finite dimensional Lie algebra E because the $H_i(E)$'s are the homology groups of a complex whose terms are the $\Lambda^i(E)$'s. The left hand side corresponds to a product over the positive roots because $\Lambda(A \oplus B) = \Lambda(A) \otimes \Lambda(B)$, $\Lambda(A) = 1 - A$ if A is one dimensional, and E is the sum of mult(α) one dimensional spaces for each positive root α. For infinite dimensional Lie algebras E we need to be careful that the infinite dimensional virtual vector spaces $H(E)$ and $\Lambda(E)$ are well defined; in this paper they are always differences of graded vector spaces with finite dimensional homogeneous pieces and so are well defined. It is more difficult to identify $H(E)$ with a sum over the Weyl group, and we do this roughly as follows. For Kac-Moody algebras $H_i(E)$ turns out to have dimension equal to the number of elements in the Weyl group of length i; for finite dimensional Lie algebras this was first observed by Bott, and was used by Kostant [26] to give a homological proof of the Weyl character formula. The sum over the homology groups can therefore be identified with a sum over the Weyl group.

For Kac-Moody algebras the same is true and was proved by Garland and Lepowsky [20]. For generalized Kac-Moody algebras things are a bit more complicated. The sum over the homology groups can still be identified with a sum over the Weyl group, but the things we sum are more complicated and contain terms corresponding to the imaginary simple roots.

We can work out the homology groups of E explicitly provided we know the simple roots of our Lie algebra M; for example, the first homology group $H_1(E)$ is the sum of the simple root spaces. For the monster Lie algebra we have worked out the simple roots using its denominator formula, which is the product formula for the j function. In section 8 we use this to work out the homology groups of E, and they turn out to be $H_0(E) = \mathbf{R}$, $H_1(E) = \sum_{n\in\mathbf{Z}} V_n pq^n$, $H_2(E) = \sum_{m>0} V_m p^{m+1}$, and all the higher homology groups are 0. Each homology group is a $\mathbf{Z}^2$-graded representation of the monster, and we use the p's and q's to keep track of the grading. If we substitute these values into the formula $\Lambda(E) = H(E)$ we find that

$$\Lambda(\sum_{n\in\mathbf{Z},m>0} V_{mn}p^m q^n) = \sum_m V_m p^{m+1} - \sum_n V_n pq^n.$$

Both sides of this are virtual graded representations of the monster. If we replace everything by its dimension we recover the product formula for the j function. More generally, we can take the trace of some element of the monster on both sides, which after some calculation gives the identity

$$p^{-1}\exp\Big(-\sum_{i>0}\sum_{m>0,n\in\mathbf{Z}} \mathrm{Tr}(g^i|V_{mn})p^{mi}q^{ni}/i\Big) = \sum_{m\in\mathbf{Z}} \mathrm{Tr}(g|V_m)p^m - \sum_{n\in\mathbf{Z}} \mathrm{Tr}(g|V_n)q^n$$

where $\mathrm{Tr}(g|V_n)$ is the trace of g on the vector space V_n.

These relations between the coefficients $\mathrm{Tr}(g|V_n)$ of the Thompson series are strong enough to determine them from their first few coefficients. Norton and Koike checked that certain modular functions of genus 0 also satisfy the same recursion relations, so we can prove that the Thompson series $T_g(q)$ are these modular functions of genus 0 by checking that the first few coefficients of both functions are the same. Unfortunately this final step of the proof (in section 9) is a case by case check that the first few coefficients are the same. Norton has conjectured [28] that Hauptmoduls with integer coefficients are essentially the same as functions satisfying relations similar to the ones above, and a conceptual proof or explanation of this would be a big improvement to the final step of the proof. (It should be possible to prove Norton's conjecture by a very long and tedious case by case check, because all functions which are either Hauptmoduls or which satisfy the relations above can be listed explicitly. In [1] the authors use a computer to find all "completely replicable" functions with integer coefficients, and they all appear to be Hauptmoduls. Roughly half of them correspond to conjugacy classes of the monster.)

I thank J. McKay, U. Tillman, J. Lepowsky, and the referee, each of whom sent me many useful remarks about a draft of this paper.

Notation. (In roughly alphabetical order.)

$\mathbf{C}$ The complex numbers.

$c(n)$ are the coefficients of the elliptic modular function $j(q) - 744$ (defined below).

$c_g(n)$ $= \mathrm{Tr}(g|V_n)$ is the n'th coefficient of the Thompson series $T_g(q)$ of g.

$\Gamma_0(N)$ is the subgroup of matrices $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$ in $SL_2(\mathbf{Z})$ with $N|c$, and $\Gamma_0(N)+$ is its normalizer in $SL_2(\mathbf{R})$; see [13].

δ_i^j is the Kronecker delta function, which is 1 if $i = j$ and 0 otherwise.

$\Delta(q)$ is the Dedekind delta function $q\prod_{n>0}(1-q^n)^{24} = \eta(q)^{24}$

E The subalgebra of a generalized Kac-Moody algebra $G = F \oplus H \oplus E$ spanned by the positive root spaces.

E_8 The unique 8-dimensional positive definite even unimodular lattice.

$\epsilon(\alpha)$ The coefficient associated with the root α in the denominator formula of s generalized Kac-Moody algebra. See section 4.

$\eta(q)$ is the Dedekind eta function $q^{1/24}\prod_{n>0}(1-q^n)$. For η_g see sections 9 or 13, and for η_+, η_- see section 11.

$\theta_\Lambda(q)$ $= \sum_{\lambda\in\Lambda} q^{\lambda^2/2} = 1 + 196560q^2 + \ldots$ is the theta function of the Leech lattice Λ. For other theta functions see section 11.

G A generalized Kac-Moody algebra; see section 4.

H The Cartan subalgebra of a generalized Kac-Moody algebra $G = F \oplus H \oplus E$.

$H_i(E)$ is a homology group of the Lie algebra E; $H(E)$ is the alternating sum of the homology groups of E.

$II_{1,1}$ is the unique even 2-dimensional unimodular Lorentzian lattice, which has inner product matrix $\left(\begin{smallmatrix} 0 & -1 \\ -1 & 0 \end{smallmatrix}\right)$. Its elements are usually represented as pairs $(m, n) \in \mathbf{Z}^2 = \mathbf{Z} \oplus \mathbf{Z}$, and this element has norm $-2mn$.

$II_{25,1}$ is the unique even 26-dimensional unimodular Lorentzian lattice, which is isomorphic to $\Lambda \oplus II_{1,1}$.

$j(q)$ is the elliptic modular function with $j(q) - 744 = q^{-1} + 196884q + \ldots = \theta_\Lambda(q)/\Delta(q) - 24 = \sum_n c(n)q^n$

Λ is the Leech lattice, the unique 24 dimensional even unimodular positive definite lattice with no vectors of norm 2. Its elements will often be denoted by λ. For its double cover $\hat{\Lambda}$ see section 12.

$\Lambda(E)$ is the alternating sum of the exterior powers of the vector space E.

$\Lambda^i(E)$ is the i'th exterior power of the vector space E.

M $= \oplus_{m,n\in\mathbf{Z}} M_{m,n}$ is the monster Lie algebra with root lattice $II_{1,1}$, constructed in section 6.

M_Λ is the fake monster Lie algebra, whose root lattice is $II_{25,1}$.

$\mu(d)$ is the Moebius function, equal to $(-1)^{(\text{number of prime factors of } d)}$ if d is square free, and 0 otherwise.

$\mathrm{mult}(r)$ The multiplicity of the root r.

p A formal variable. It can usually be considered as a complex number with $|p| < 1$.

$p_{24}(n)$ is the number of partitions of n into parts of 24 colours, so that $\sum_n p_{24}(1+n)q^n = q^{-1}\prod_{n>0}(1-q^n)^{-24} = \Delta(q)^{-1} = q^{-1} + 24 + 324q + 3200q^2 + 25650q^3 + 176256q^4 + 1073720q^5 + \ldots$. These are the multiplicities of roots of the fake monster Lie algebra M_Λ. For p_g see section 14.

q A formal variable. It can usually be thought of as a complex number with $|q| < 1$, equal to $e^{2\pi i\tau}$. (I.e., the formal series usually converge for $|q| < 1$.)

r^2 The norm (r, r) of the vector r of some lattice.

$\mathbf{R}$ The real numbers.

ρ is the Weyl vector of a root lattice, which by definition has the property that $(\rho, r) = -(r, r)/2$ for any simple root r. (The Weyl vector is not necessarily in the root lattice, although it does for most of the Lie algebras in this paper.) This has the opposite sign to the usual convention for the Weyl vector, for reasons explained in section 4. For the root lattice $II_{25,1} = \Lambda \oplus II_{1,1} = \Lambda \oplus \mathbf{Z}^2$ of the fake monster Lie algebra M_Λ, ρ is the vector $(0, 0, 1)$.

$\sigma_i(n)$ $= \sum_{d|n} d^i$ is the sum of the i'th powers of the divisors of n.

$T_g(q)$ $= \sum c_g(n)q^n$ is the Thompson series of an element g of the monster, with $c_g(n) = \mathrm{Tr}(g|V_n)$ where $V = \oplus_{n\in\mathbf{Z}} V_n$ is the module constructed by Frenkel, Lepowsky and Meurman [16,17]. (The spaces V and V_n are denoted by $V^\natural$ and $V^\natural_{-n}$ in [16].)

$\mathrm{Tr}(g|U)$ is the trace of an endomorphism g of a vector space U.

τ A complex number with $\mathrm{Im}(\tau) > 0$.

V $= \oplus_{n\in\mathbf{Z}} V_n$ is the monster vertex algebra discussed in section 3.

$V_{II_{1,1}}$ is the vertex algebra of the two dimensional even Lorentzian lattice $II_{1,1}$ (or more precisely the vertex algebra of its double cover).

V_Λ is the the fake monster vertex algebra, which is the vertex algebra of the Leech lattice Λ (or more precisely of its double cover $\hat{\Lambda}$). See section 12.

W is a Weyl group. Typical elements are often denoted by w. See section 4.

$\mathbf{Z}$ The integers.

ψ^i An Adams operation on virtual group representations, defined by $\mathrm{Tr}(g|\psi^i(V)) = \mathrm{Tr}(g^i|V)$ for V a virtual representation of a group containing g.

ω A Cartan involution (section 4) or a conformal vector of a vertex algebra (section 3).

2 Introduction (continued).

We describe the results in the second half of the paper (sections 10 to 14). This section can be omitted by those who are only interested in the proof of theorem 1.1.

We construct several Lie superalgebras which are similar to the monster Lie algebra. One method of constructing these is to consider the "twisted" denominator formulas of the monster Lie algebra as untwisted denominator formulas of some other Lie algebras or superalgebras; these seem to be related to other sporadic simple groups. A second method of constructing some of them is to replace the monster vertex algebra V by the vertex algebra of the Leech lattice V_Λ. From this we get the fake monster Lie algebra [8] (where it is called the monster Lie algebra) and several variations of it.

The Lie superalgebras we construct form two families as follows:

(1) A Lie algebra or superalgebra of rank 2 for many conjugacy classes g of the monster simple group. The monster Lie algebra is the one corresponding to the identity element of the monster. The ones corresponding to other elements of the monster are often related to other sporadic simple groups.

(2) A Lie superalgebra for many of the conjugacy classes of the group $\mathrm{Aut}(\hat{\Lambda}) = 2^{24}.2.Co_1$, where $\hat{\Lambda}$ is the standard double cover of the Leech lattice Λ (defined in section 12),

$\mathrm{Aut}(\hat{\Lambda})$ is the group of its automorphisms which preserve the inner product on Λ, and $Co_1 = \mathrm{Aut}(\Lambda)/\mathbf{Z}_2$ is one of Conway's sporadic simple groups. (The symbol $A.B$ where A and B are groups stands for some extension of B by A, i.e., a group with a normal subgroup A such that the quotient by A is B; the notation is ambiguous.) For example, we get Lie algebras of ranks 26, 18, and 14 corresponding to certain automorphisms of $\hat{\Lambda}$ of orders 1, 2, and 3, and a Lie superalgebra of rank 10 corresponding to an automorphism of order 2. The Lie algebra of rank 26 is what we now call the fake monster Lie algebra and is studied in [8] (where it is called the monster Lie algebra, because the genuine monster Lie algebra had not been discovered then).

All of these algebras are generalized Kac-Moody algebras or superalgebras. Their root multiplicities can be described explicitly in terms of the coefficients of a finite number of modular forms of weight at most 0. (These modular forms are holomorphic on the upper half plane, and are meromorphic but not necessarily holomorphic at the cusps.) More precisely, there is a sublattice L of finite index in the root lattice, and for each coset of L in the root lattice there is a modular form, which is holomorphic except at the cusps and of weight $1 - \dim(L)/2$, such that the root multiplicity of a vector r in a given coset is the coefficient of $q^{-(r,r)/2}$ of the corresponding modular form. All the algebras have Weyl vectors ρ with norm $\rho^2 = (\rho,\rho) = 0$, and the simple roots are the roots r with $(r,\rho) = -(r,r)/2$. In particular the simple roots and root multiplicities can be described explicitly, and we use this to obtain some infinite product identities from the denominator formulas of these algebras. These algebras are closely related to the sporadic simple groups.

We now discuss both of these families of Lie algebras in more detail. We recall from section 1 that the monster Lie algebra M, is a generalized Kac-Moody algebra whose denominator formula is

$$p^{-1}\prod_{\substack{m>0\\ n\in\mathbf{Z}}}(1-p^m q^n)^{c(mn)} = j(p) - j(q). \tag{2.1}$$

In section 10 we construct a similar Lie superalgebra for many elements g of the monster. In this case the denominator formula is

$$p^{-1}\prod_{\substack{m>0\\ n\in\mathbf{Z}}}(1-p^m q^n)^{\mathrm{mult}(m,n)} = T_g(p) - T_g(q) \tag{2.2}$$

where $T_g(q) = \sum \mathrm{Tr}(g|V_n)q^n$ is the Thompson series of the element g, and the multiplicity $\mathrm{mult}(m,n)$ of the root (m,n) is given by

$$\mathrm{mult}(m,n) = \sum_{ds|(m,n)} \mu(d)\mathrm{Tr}(g^s|V_{mn/d^2s^2})/ds \tag{2.3}$$

where $\mu(d)$ is the Moebius function, equal to $(-1)^{(\text{number of prime factors of } d)}$ if d is square free, and 0 otherwise. (The symbol (m,n) under the summation sign means the greatest common divisor of m and n, rather than the ordered pair.) For example, if N

is a squarefree integer such that the full normalizer $\Gamma_0(N)+$ of $\Gamma_0(N)$ has genus 0 then a special case of (2.2) is

$$p^{-1}\prod_{\substack{m>0\\ n\in\mathbf{Z}}}\prod_{d|(m,n,N)}(1-p^mq^n)^{c_N(mn/d)}=T_N(p)-T_N(q) \tag{2.4}$$

where $T_N(q)=\sum_n c_N(n)q^n$ is the normalized generator of the function field of $\Gamma_0(N)+$ with leading terms $q^{-1}+0+c_N(1)q+\dots$.

In section 11 to 14 we construct a second series of Lie superalgebras, which are similar to the algebras above except that they are related to the vertex algebra V_Λ of the Leech lattice Λ instead of the monster vertex algebra V. The largest of these, which plays the same role for this series as the monster Lie algebra plays for the previous series, is the algebra which used to be called the monster Lie algebra in [8] and is now called the fake monster Lie algebra. (The Kac-Moody algebra whose Dynkin diagram is that of the reflection group of $II_{25,1}$ has also been called the monster Lie algebra ([7]); it is a large subalgebra of the fake monster Lie algebra, and does not seem to be interesting, except as an approximation to the fake monster Lie algebra.) The root lattice of the fake monster Lie algebra is the 26 dimensional even unimodular Lorentzian lattice $II_{25,1}$, and its denominator formula is

$$e^\rho\prod_{r\in\Pi^+}(1-e^r)^{p_{24}(1-r^2/2)}=\sum_{w\in W}\det(w)w(e^\rho\prod_{n>0}(1-e^{n\rho})^{24}).$$

Here ρ is a norm 0 Weyl vector for the reflection group W of $II_{25,1}$, Π^+ is the set of positive roots, which is the set of vectors r of norm at most 2 which are either positive multiples of ρ or have negative inner product with ρ, $p_{24}(1-r^2/2)$ is the multiplicity of the root r and is equal to the number of partitions of the integer $1-(r,r)/2$ into parts of 24 colours, and the simple roots are the norm 2 vectors r with $(r,\rho)=-1$ together with 24 copies of each positive multiple of ρ. This Lie algebra was first constructed in [3], and the properties stated above were proved in [8]; this construction depended heavily on the ideas in Frenkel [15].

The fake monster Lie algebra is acted on by the group $2^{24}.2.Co_1$ in the same way that the monster Lie algebra is acted on by the monster group, and we construct a superalgebra for many elements of $2^{24}.2.Co_1$ from the fake monster Lie algebra in the same way that we construct a superalgebra for every element of the monster from the monster Lie algebra. Some of the more interesting algebras we get in this way are a fake baby monster algebra of rank 18, a fake Fi_{24} Lie algebra of rank 14 associated with Fischer's sporadic simple group Fi_{24}, a fake Co_1 superalgebra of rank 10 associated with Conway's sporadic simple group Co_1, and several Lie algebras of smaller rank corresponding to some of the other sporadic simple groups involved in the monster.

The superalgebra of rank 10 is particularly interesting, so we describe it explicitly. It is the superalgebra associated with an element g of $\mathrm{Aut}(\hat\Lambda)$ which has order 2 and fixes an 8-dimensional sublattice of Λ, isomorphic to the lattice E_8 with all norms doubled. Its root lattice L is the dual of the sublattice of even vectors of $I_{9,1}$, so that L is a nonintegral lattice of determinant 1/4 all of whose vectors have integral norm. We represent vectors

of L as triples (v, m, n), where $v \in E_8$ and $m, n \in \mathbf{Z}$, and (v, m, n) has norm $v^2 - 2mn$. The lattice $I_{9,1}$ is then the set of vectors (v, m, n) with $m + n$ even. We let ρ be the norm 0 vector $(0, 0, 1)$. We let the Weyl group W be the subgroup of $\mathrm{Aut}(L)$ generated by the reflections of norm 1 vectors, so that the simple roots of W are the norm 1 vectors r with $(r, \rho) = -1/2$. (Note that W is not the full reflection group of L, as L also has roots of norm 2.)

The simple roots of our superalgebra are the simple roots of W, together with the positive multiples $n\rho$ $(n > 0)$ of ρ, each with multiplicity $8(-1)^n$. Here we use the convention that multiplicity $-k < 0$ means a superroot of multiplicity k, so that the odd multiples of ρ are superroots. In general the vector (v, m, n) is an ordinary root or a superroot depending on whether $m + n$ is even or odd, so the ordinary roots are those in the sublattice $I_{9,1}$ of L.

The multiplicity of the root $r = (v, m, n) \in L$ is equal to

$$\mathrm{mult}(r) = (-1)^{(m-1)(n-1)} p_g((1 - r^2)/2) = (-1)^{m+n} |p_g((1 - r^2)/2)|,$$

where $p_g(n)$ is defined by

$$\sum p_g(n) q^n = q^{-1/2} \prod_{n>0} (1 - q^{n/2})^{-(-1)^n 8}$$

and as before negative multiplicity means a superroot. The denominator formula for this Lie superalgebra is

$$e^\rho \prod_{r \in \Pi^+} (1 - e^r)^{\mathrm{mult}(r)} = \sum_{w \in W} \det(w) w(e^\rho \prod_{n>0} (1 - e^{n\rho})^{(-1)^n 8}).$$

Similarly the denominator formula for the fake baby monster Lie algebra of rank 18 is

$$e^\rho \prod_{r \in L^+} (1 - e^r)^{p_g(1 - r^2/2)} \prod_{r \in 2L'^+} (1 - e^r)^{p_g(1 - r^2/4)} = \sum_{w \in W} \det(w) w(e^\rho \prod_{i>0} (1 - e^{i\rho})^8 (1 - e^{2i\rho})^8)$$

where L is the Lorentzian lattice which is the sum of the 16-dimensional Barnes-Wall lattice Λ_2 [12] and the two dimensional even Lorentzian lattice $II_{1,1}$, W is its reflection group which has Weyl vector ρ, L' is the dual of L and p_g is defined by $\sum_n p_g(n) q^n = \prod_{n>0} (1 - q^n)^{-8} (1 - q^{2n})^{-8}$. This Lie algebra is associated with an element g of order 2 of $\mathrm{Aut}(\hat{\Lambda})$ which fixes a 16-dimensional lattice of Λ isomorphic to the Barnes-Wall lattice.

There are similar Lie superalgebras associated with many other elements of $\mathrm{Aut}(\hat{\Lambda})$, which have denominator formulas similar to those above.

3 Vertex algebras.

We give a brief summary of the facts about vertex algebras that we use and list the properties of the monster vertex algebra that we need. For more information about vertex

algebras, see [3], [9], [16], or [18]; the last two references contain proofs of the results quoted here.

A vertex algebra over the real numbers (which is the only case we use in this paper) is a vector space V over the real numbers with an infinite number of bilinear products, written $u_n v$ for $u, v \in V, n \in \mathbf{Z}$, such that

(1) $u_n v = 0$ for n sufficiently large (depending on u and v).

(2)

$$\sum_{i \in \mathbf{Z}} \binom{m}{i} (u_{q+i}v)_{m+n-i}w = \sum_{i \in \mathbf{Z}} (-1)^i \binom{q}{i} (u_{m+q-i}(v_{n+i}w) - (-1)^q v_{n+q-i}(u_{m+i}w))$$

for all u, v, and w in V and all integers m, n and q.

(3) There is an element $1 \in V$ such that $v_n 1 = 0$ if $n \geq 0$ and $v_{-1}1 = v$.

We often think of u_n as a linear map from V to V, taking v to $u_n v$. In [16,18] these operators are combined into the "vertex operator" $V(u,z) = \sum_{n \in \mathbf{Z}} u_n z^{-n-1}$, which is an operator valued formal Laurent series in z.

For example, if V is a commutative algebra over $\mathbf{R}$ with derivation D then we can make it into a vertex algebra by defining $u_n v$ to be $D^{-n-1}(u)v/(-n-1)!$ for $n < 0$ and 0 for $n \geq 0$, and conversely any vertex algebra over $\mathbf{R}$ for which $u_n v = 0$ whenever $n \geq 0$ arises from a commutative algebra in this way. These are the only finite dimensional examples of vertex algebras, and seem to be the only examples which are easy to construct. The vertex algebras we use here do not have this property and are much harder to construct; for example the detailed construction of the monster vertex algebra in [16] takes up most of a rather long book. Fortunately the details of this construction are not important for this paper, so at the end of this section we list the properties of it that we need.

We define the operator D of a vertex algebra by $D(v) = v_{-2}1$. (In [16] and [18] D is denoted by L_{-1}.) The vector space V/DV is a Lie algebra, where the bracket is defined by $[u, v] = u_0 v$ and DV is the image of V under D.

There is a vertex algebra $V_{\hat{L}}$ associated with any even lattice L (or more precisely with a certain central extension $\hat{L}$ of L by a group of order 2), which is constructed in [3]. The underlying vector space of this vertex algebra is the tensor product of the twisted group ring $Q(\hat{L})$ of the double cover $\hat{L}$ of L (see section 12) and the ring of polynomials $S(\oplus_{i>0} L_i)$ over the sum of an infinite number of copies L_i of $L \otimes \mathbf{R}$. We often use the vertex algebra $V_{II_{1,1}}$ of the 2-dimensional even unimodular Lorentzian lattice $II_{1,1}$. There is also a monster vertex algebra V ([3], [16]) which is acted on naturally by the monster sporadic simple group. (The referee has asked me to explain why I have never published my (long and messy) proof of the assertion in [3] that the module V constructed in [17] has the structure of a vertex algebra. The reason is that my proof used many results from the announcement [17], and the only published proof of this announcement, given in the book [16], incorporates the fact that V is a vertex algebra.) The underlying vector space of the monster vertex algebra is a graded vector space $V = \oplus_{n \in \mathbf{Z}} V_n$, with homogeneous pieces of dimensions 1, 0, 196884, ... equal to the coefficients of the elliptic modular function $j(q) - 744$.

A conformal vector of dimension (or "central charge") $c \in \mathbf{R}$ of a vertex algebra V is defined to be an element ω of V such that $\omega_0 v = D(v)$ for any $v \in V$, $\omega_1 \omega = 2\omega$,

$\omega_3\omega = c/2$, $\omega_i\omega = 0$ if $i = 2$ or $i > 3$, and any element of V is a sum of eigenvectors of the operator $L_0 = \omega_1$ with integral eigenvalues. If v is an eigenvector of L_0, then its eigenvalue is called the (conformal) weight of v. If v is an element of the monster vertex algebra V of conformal weight n, we sometimes say that v has degree $n - 1$ $(= n - c/24)$.

The vertex algebra of any c-dimensional even lattice has a canonical conformal vector of dimension c, and the monster vertex algebra has a conformal vector of dimension 24. If ω is a conformal vector of a vertex algebra V then we define operators L_i on V for $i \in \mathbf{Z}$ by

$$L_i = \omega_{i+1}.$$

These operators satisfy the relations

$$[L_i, L_j] = (i - j)L_{i+j} + \binom{i+1}{3}\frac{c}{2}\delta^i_{-j}$$

and so make V into a module over the Virasoro algebra. The operator L_{-1} is equal to D. We define the space P^i to be the space of vectors $w \in V$ such that $L_0(w) = iw$, $L_i(w) = 0$ if $i > 0$. The space $P^1/(DV \cap P^1)$ is a subalgebra of the Lie algebra V/DV, which is equal to P^1/DP^0 for the vertex algebras we use in this paper.

The vertex algebra of any even lattice and the monster vertex algebra both have a real valued symmetric bilinear form (,) such that the adjoint of the operator u_n is $(-1)^i \sum_{j\geq 0} L_1^j(\omega(u))_{2i-j-n-2}/j!$ if u has degree i, where ω is the automorphism of the vertex algebra defined by $\omega(e^w) = (-1)^{(w,w)/2}(e^w)^{-1}$ for e^w an element of the twisted group ring of L corresponding to the vector $w \in L$, or is 1 in the case of the monster vertex algebra. (There is an unfortunate clash of notation here, because ω is the standard notation for both the conformal vector and the Cartan automorphism of a Lie algebra. If L is the root lattice of a finite dimensional Lie algebra of type A_n, D_n, E_6, E_7, or E_8 then the Lie algebra is a subalgebra of the vertex algebra of L and the involution ω defined above restricts to the Cartan automorphism of the Lie algebra.) If a vertex algebra has a bilinear form with the properties above we say that the bilinear form is compatible with the conformal vector.

Frenkel, Lepowsky, and Meurman [16] use "vertex operator algebras", rather than vertex algebras, so we explain what the difference is. A vertex operator algebra is a vertex algebra with a conformal vector such that the eigenspaces of L_0 are all finite dimensional with nonnegative integral eigenvalues. For example, the monster vertex algebra is a vertex operator algebra. (Its conformal vector spans the subspace of V_1 fixed by the monster.) The vertex algebras used in this paper all have conformal vectors but the eigenspaces of L_0 are not always finite dimensional. An example of a vertex algebra without these properties is the vertex algebra $V_{II_{1,1}}$ of the 2-dimensional even unimodular Lorentzian lattice $II_{1,1}$.

If V and W are vertex algebras, then their tensor product $V \otimes W$ as vector spaces is also a vertex algebra, if $(a \otimes b)_n(c \otimes d)$ is defined to be $\sum_{i\in\mathbf{Z}}(a_i c) \otimes (b_{n-1-i}d)$ and the identity 1 is defined to be $1 \otimes 1$. (See [18].) If V and W have conformal vectors ω_V and ω_W of dimensions m and n then $\omega_V \otimes 1 + 1 \otimes \omega_W$ is a conformal vector of $V \otimes W$ of dimension $m + n$. If V and W have bilinear forms compatible with the conformal vectors then so does $V \otimes W$; in fact the obvious bilinear form on the tensor product will do.

The monster vertex algebra V [3, 16] is a vertex algebra acted on by the monster. The construction given in [16] is rather long, but fortunately we do not need to know many details of its construction. We list the three properties of it we do use, in case anyone finds a simpler construction of it that they wish to prove also satisfies Conway and Norton's moonshine conjectures.

(1) V is a vertex algebra over $\mathbf{R}$ with a conformal vector ω of dimension 24 and a positive definite bilinear form such that the adjoint of u_n is given by the expression above (with the automorphism ω acting trivially).

(2) V is a sum of eigenspaces V_i of the operator L_0, where V_i is the eigenspace on which L_0 has eigenvalue $i+1$, and the dimension of V_i is given by $\sum \dim(V_i)q^i = j(q) - 744 = q^{-1} + 196884q + \ldots$.

These two conditions turn out to imply that if g is any automorphism of V preserving the vertex algebra structure, the conformal vector and the bilinear form then $\sum \mathrm{Tr}(g|V_i)q^i$ is a completely replicable function (which means that identity (8.3) holds).

(3) The monster simple group acts on V, preserving the vertex algebra structure, the conformal vector ω and the bilinear form. The first few representations V_i of the monster (after $V_{-1} = \chi_1$, $V_0 = 0$) decompose as $V_1 = \chi_1 + \chi_2$, $V_2 = \chi_1 + \chi_2 + \chi_3$, $V_3 = 2\chi_1 + 2\chi_2 + \chi_3 + \chi_4$, $V_5 = 4\chi_1 + 5\chi_2 + 3\chi_3 + 2\chi_4 + \chi_5 + \chi_6 + \chi_7$ where χ_i, $1 \le i \le 7$ are the first seven irreducible representations of the monster, indexed in order of increasing dimension.

The construction of the vector space V and the action of the monster on it, preserving a small part of the vertex algebra structure, were announced in [17], and the vertex algebra structure was announced in [3]. The results quoted above have all been proved in [16] and [18], apart from the explicit description of the representations in (3). We show in section 9 that condition (3) can be proved using the explicit description in [16] of the traces of certain elements of the monster acting on V. In the course of proving theorem 1.1 we show that these three conditions characterize V as a graded representation of the monster.

4 Generalized Kac-Moody algebras

We summarize the results about generalized Kac-Moody algebras that we use, which can be found in [4], [5], [20], and the third edition of Kac's book [23]. We modify the original definition of generalized Kac-Moody algebras in [4] slightly so that these algebras are closed under taking universal central extensions, as in [5]. (This is not necessary for the proof of theorem 1.1, but makes the converse to theorem 4.1 slightly neater.) All Lie algebras are Lie algebras over the reals.

A Lie algebra G is defined to be a (split) generalized Kac-Moody algebra if it has an almost positive definite contravariant bilinear form, which means that G has the following three properties.

1. G can be $\mathbf{Z}$-graded as $G = \oplus_{i \in \mathbf{Z}} G_i$ and G_i is finite dimensional if $i \ne 0$.
2. G has an involution ω which maps G_i into G_{-i} and acts as -1 on G_0, so in particular G_0 is abelian.
3. G has an invariant bilinear form $(,)$ invariant under ω such that G_i and G_j are orthogonal if $i \ne -j$, and such that $-(g, \omega(g)) > 0$ if g is a nonzero homogeneous element of G of nonzero degree.

The bilinear form $(,)_0$ defined by $(a,b)_0 = -(a,\omega(b))$ is called the contravariant bilinear form of G, and is positive definite on G_i if $i \neq 0$. For example any finite dimensional split semisimple Lie algebra over the reals has these properties, with ω equal to a Cartan involution, $(,)$ equal to the Killing form, and the grading determined by the eigenvalues of some suitably normalized regular element. (If the contravariant bilinear form on G is positive semidefinite, then G is essentially a sum of affine, Heisenberg, and finite dimensional simple split Lie algebras.) In general the contravariant form can be indefinite on G_0. For the algebras we use in this paper the subalgebra G_0 has a vector such that the contravariant form is positive semidefinite on its orthogonal complement.

Suppose that $a_{ij}, i, j \in I$ is a symmetric countable (possibly infinite) real matrix such that $a_{ij} \leq 0$ if $i \neq j$ and such that if $a_{ii} > 0$ then $2a_{ij}/a_{ii}$ is an integer for any j. Then the universal generalized Kac-Moody algebra G of this matrix is defined to be the Lie algebra generated by elements e_i, f_i, h_{ij} for $i, j \in I$ satisfying the following relations. (It is true but not obvious that any universal generalized Kac-Moody algebra is a generalized Kac-Moody algebra; this fact is not needed in this paper.)

1. $[e_i, f_j] = h_{ij}$
2. $[h_{ij}, e_k] = \delta_i^j a_{ik} e_k, [h_{ij}, f_k] = -\delta_i^j a_{ik} f_k$
3. If $a_{ii} > 0$ and $i \neq j$ then $\mathrm{ad}(e_i)^n e_j = \mathrm{ad}(f_i)^n f_j = 0$, where $n = 1 - 2a_{ij}/a_{ii}$.
4. If $a_{ii} \leq 0, a_{jj} \leq 0$ and $a_{ij} = 0$ then $[e_i, e_j] = [f_i, f_j] = 0$.

We often write h_i for h_{ii}. There is a unique invariant bilinear form on this algebra such that $(e_i, f_j) = \delta_i^j$; this implies that $(h_i, h_j) = a_{ij}$.

Remark. If $a_{ii} > 0$ for all $i \in I$ then this algebra is the same as the Kac-Moody algebra with symmetrized Cartan matrix a_{ij}. In general these algebras have almost all the properties that Kac-Moody algebras have, and the only major difference is that generalized Kac-Moody algebras are allowed to have imaginary simple roots.

The main theorem about generalized Kac-Moody algebras [4,5] says that they can all be obtained from universal generalized Kac-Moody algebras.

Theorem 4.1. *Suppose that G is a generalized Kac-Moody algebra with some given grading, involution ω and bilinear form $(,)$. Then there is a unique universal generalized Kac-Moody algebra, graded by putting $\deg(e_i) = -\deg(f_i) = n_i$ for some positive integers n_i, with a homomorphism f (not necessarily unique) to G such that*

(1) f preserves the gradings, involutions and bilinear forms (where the universal generalized Kac-Moody algebra is given the grading, involution and bilinear form defined above).

(2) The kernel of f is in the centre of the universal generalized Kac-Moody algebra (which is contained in the abelian subalgebra spanned by the elements h_{ij}).

(3) The image of f is an ideal of G, and G is the semidirect product of this subalgebra and a subalgebra of the abelian subalgebra G_0. Moreover the images of all the generators e_i and f_i are eigenvectors of G_0.

Remark. The converse of this theorem is also true (but not needed here): any universal generalized Kac-Moody algebra with a countable number of simple roots is a generalized Kac-Moody algebra. (The proof of this converse in [4] contains a gap which was pointed out and corrected by Kac in [23].) Theorem 4.1 says that we can construct any generalized

Kac-Moody algebra from some universal generalized Kac-Moody algebra by factoring out some of the centre and adding a commuting algebra of outer derivations.

We list some of the properties of universal generalized Kac-Moody algebras G from [4] and [5], many of which are also proved in the third edition of [23].

(1) The element h_{ij} is 0 unless the i'th and j'th columns of a are equal. The elements h_{ij} for which the i'th and j'th columns of a are equal form a basis for an abelian subalgebra H of G, called its Cartan subalgebra. In the case of Kac-Moody algebras, the i'th and j'th columns of a cannot be equal unless $i = j$, so the only nonzero elements h_{ij} are those of the form h_{ii}, which are usually denoted by h_i. Every nonzero ideal of G has nonzero intersection with H. The centre of G is contained in H and contains all the elements h_{ij} for $i \neq j$.

(2) The root lattice is defined to be the free abelian group generated by elements r_i for $i \in I$, with the bilinear form given by $(r_i, r_j) = a_{ij}$. The elements r_i are called the simple roots. The universal generalized Kac-Moody algebra is graded by the root lattice if we let e_i have degree r_i and f_i have degree $-r_i$. There is a natural homomorphism of abelian groups from the root lattice to the Cartan subalgebra H taking r_i to h_i which preserves the bilinear forms. If r is in the root lattice then the vector space of elements of the Lie algebra of that degree is called the root space of r; if r is nonzero and has a nonzero root space then r is called a root of the generalized Kac-Moody algebra. A root r is called positive if it is a sum of simple roots, and negative otherwise. It is called real if it has positive norm (r, r) and imaginary otherwise.

(3) The Weyl vector ρ is the additive map from the root lattice to $\mathbf{R}$ taking r_i to $-(r_i, r_i)/2$ for all $i \in I$. (This is the opposite sign from the usual definition.) We write its value on a root r as (ρ, r). The fundamental Weyl chamber (which we call the Weyl chamber for short, because it is the only one we use) is the set of vectors v of the Cartan subalgebra H with $(v, r_i) \leq 0$ for all real simple roots r_i, where we identify real simple roots with their images in H. (There is a misprint in the definition of Weyl chamber in [4]; the word "real" in "real simple roots" was omitted.)

Remark. The reason why we use nonstandard sign conventions for the Weyl vector and Weyl chamber is that we are using the convention that "everything should be in the Weyl chamber if possible". We want the imaginary simple roots to be in the Weyl chamber which forces us to use the opposite Weyl chamber from usual; similarly we change the sign of the Weyl vector to get it (or rather its image in H, when this exists) in the Weyl chamber, and we use lowest weight modules (rather than highest weight modules) because their lowest weights are in the Weyl chamber. For most of the generalized Kac-Moody algebras in this paper the Cartan subalgebra has a Lorentzian bilinear form, so that the vectors of norm at most 0 form two solid cones; our sign conventions usually make all the interesting vectors of norm at most 0 lie in the same solid cone.

(4) If a has no zero columns then G is perfect and equal to its own universal central extension. If a is not the direct sum of two smaller matrices, not the 1×1 zero matrix, and not the matrix of an affine Kac-Moody algebra, then G modulo its centre is simple.

(5) Suppose we choose a positive integer n_i for each $i \in I$. Then we can grade G by putting $\deg(e_i) = -\deg(f_i) = n_i$. The degree 0 piece of G is the Cartan subalgebra H.

(6) G has an involution ω with $\omega(e_i) = -f_i, \omega(f_i) = -e_i$, called the Cartan involution.

There is a unique invariant bilinear form (,) on G such that $(e_i, f_i) = 1$ for all i, and it also has the property that $-(g, \omega(g)) > 0$ whenever g is a homogeneous element of nonzero degree. In particular the universal generalized Kac-Moody algebra of any matrix as above is a generalized Kac-Moody algebra.

(7) There is a character formula for the simple lowest weight module M_λ of G with a lowest weight λ in the real vector space of the root lattice such that $(\lambda, r_i) \leq 0$ for $i \in I$ and $2(\lambda, r_i)/(r_i, r_i)$ is integral if the simple root r_i is real. This states that

$$Ch(M_\lambda)e^\rho \prod_{\alpha>0}(1-e^\alpha)^{\mathrm{mult}(\alpha)} = \sum_{w\in W} \det(w) w(e^\rho \sum_\alpha \epsilon(\alpha) e^{\alpha+\lambda})$$

The only case of this we need is for the trivial one dimensional module with $\lambda = 0$, when it becomes the denominator formula

$$e^\rho \prod_{\alpha>0}(1-e^\alpha)^{\mathrm{mult}(\alpha)} = \sum_{w\in W} \det(w) w(e^\rho \sum_\alpha \epsilon(\alpha) e^{\alpha})$$

Here ρ is the Weyl vector, $\alpha > 0$ means α is a positive root, and W is the Weyl group, which is the group of isometries of the root lattice generated by the reflections corresponding to the real simple roots. If $w \in W$ then $\det(w)$ is defined to be $+1$ or -1 depending on whether w is the product of an even or odd number of reflections; if the root lattice is finite dimensional this is just the usual determinant of w acting on it. We define $\epsilon(\alpha)$ for α in the root lattice to be $(-1)^n$ if α is the sum of a set of n pairwise orthogonal imaginary simple roots that are all orthogonal to λ, and 0 otherwise. If G is a Kac-Moody algebra then there are no imaginary simple roots so the sum over α is 1 and we recover the usual character and denominator formulas.

If we have a generalized Kac-Moody algebra G then we choose some universal generalized Kac-Moody algebra mapping to G as in theorem 4.1, and call the Weyl group, root lattice, and so on of the universal generalized Kac-Moody algebra the Weyl group, root lattice and so on of G.

For the generalized Kac-Moody algebras in this paper, the Cartan subalgebra H is usually finite dimensional and its bilinear form is nonsingular. We often think of roots of G as vectors of H by using the natural map from the root lattice to H, taking r_i to h_i. We will often abuse terminology and call the corresponding vectors of H roots. The map from the set of roots to H is not usually injective, so we need to make a few comments to clarify what happens. It is possible for $n > 1$ imaginary simple roots to have the same image r in H; when this happens we say that r is a simple root of multiplicity n. (In general it is possible for the same vector of H to be the image of simple and nonsimple roots, but this does not happen for the Lie algebras in this paper.) The simple roots in H are usually linearly dependent so in general there is no reason why there should exist a Weyl vector ρ in H such that $(\rho, r_i) = -r_i^2/2$ for all real simple roots r_i; however such a vector ρ does exist for the Lie algebras in this paper.

The denominator formula also needs an obvious modification if we think of the roots as elements of H. If a simple root has multiplicity $n > 1$ then the "set" of simple roots should include n copies of it, i.e., the "set" of simple roots is really a "multiset" rather

than a set. (A multiset is a "set which may contain several copies of the same object", or more precisely a map from a set to the positive integers.) The expression $\epsilon(\alpha)$ in the character formula is then the sum of terms $(-1)^n$ for all ways of writing α as a sum of a "set" (or more precisely a multiset) of pairwise orthogonal imaginary simple roots that are all orthogonal to λ. For example, if there is only one simple root r of norm 0 and multiplicity n, then $\epsilon(ir) = \binom{n}{i}(-1)^i$ because there are $\binom{n}{i}$ ways of writing ir as a sum of a multiset of pairwise orthogonal simple roots.

Remark. Most of these results can be extended to superalgebras with little difficulty, provided that we interpret roots of negative multiplicity as being "superroots". The only extra detail worth pointing out is that we have to add the condition that all simple superroots are imaginary (and not real). For example, many of the finite dimensional superalgebras are generalized Kac-Moody superalgebras with simple superroots of norm 0. Warning: The Cartan matrix of a finite dimensional superalgebra may depend on the **Z**-grading chosen.

Any generalized Kac-Moody algebra can be written as a direct sum $E \oplus H \oplus F$ of subalgebras, where H is the Cartan subalgebra and E and F are the subalgebras corresponding to the positive and negative roots. The standard sequence

$$\ldots \to \Lambda^2(E) \to \Lambda^1(E) \to \Lambda^0(E) \to 0$$

whose homology groups are those of the Lie algebra E [10] shows that $\Lambda(E) = H(E)$, where $\Lambda(E)$ is the virtual vector space $\Lambda^0(E) \ominus \Lambda^1(E) \oplus \Lambda^2(E) \ldots$ formed from the alternating sum of the exterior powers of E, and $H(E)$ is the virtual vector space $H_0(E) \ominus H_1(E) \oplus H_2(E) \ldots$ formed from the alternating sum of the homology groups of E with real coefficients. If L is the root lattice of M, then both sides are L-graded virtual vector spaces whose homogeneous pieces are finite dimensional, so these infinite sums are meaningful.

The denominator formula for any generalized Kac-Moody algebra follows from the fact that $\Lambda(E) = H(E)$ as virtual L-graded vector spaces. The character of $\Lambda(E)$ is a product over positive roots. In the case of ordinary Kac-Moody algebras the vector space $H_i(E)$ has a basis corresponding to the elements of the Weyl group of length i, so the character of $H(E)$ is a sum over the Weyl group. If we work out everything explicitly we obtain the denominator formula

$$e^\rho \prod_{\alpha>0} (1 - e^\alpha)^{\mathrm{mult}(\alpha)} = \sum_{w \in W} \det(w) w(e^\rho)$$

for Kac-Moody algebras, and a similar but slightly more complicated calculation gives the denominator formula for generalized Kac-Moody algebras. For Kac-Moody algebras this calculation is carried out in Garland and Lepowsky [20] and most of their methods and results apply to generalized Kac-Moody algebras with only minor changes. They obtain a more general result which not only gives the character formula of any highest weight module, but also works if E is replaced by various other radicals of parabolic subalgebras. We do not need these more general results, but they are easily extended to the case of generalized Kac-Moody algebras.

The groups $H_i(E)$ can be calculated for generalized Kac-Moody algebras in the same way that they are calculated for ordinary Kac-Moody algebras in Garland and Lepowsky

[20]. The result we get is that $H_i(E)$ is the subspace of $\Lambda^i(E)$ spanned by the homogeneous vectors of $\Lambda^i(E)$ whose degrees $r \in L$ satisfy $(r+\rho)^2 = \rho^2$. The degree r of any homogeneous vector of $\Lambda^i(E)$ satisfies $(r+\rho)^2 \leq \rho^2$, so $H_i(E)$ can be thought of as a sort of boundary of $\Lambda^i(E)$. These homology groups are easy to work out for any generalized Kac-Moody algebra, provided that we know the simple roots and their multiplicities. For the monster Lie algebra, we carry out this calculation in section 8.

In practice it is sufficient to know the subspace of $H(E)$ of elements whose degree r has the property that $r+\rho$ is in the fundamental Weyl chamber. This subspace is isomorphic to the subspace of $\Lambda(E)$ of all elements that can be written in the form $e_1 \wedge e_2 \wedge \dots$ where the e_i's are vectors in the root spaces of pairwise orthogonal imaginary simple roots, so that the character of this subspace is just the sum $\sum_\alpha \epsilon(\alpha) e^\alpha$ that appears in the denominator formula. There is a similar description of the homology groups of E with coefficients in any lowest weight module which can be used to prove the character formula for lowest weight modules, but we do not use this.

5 The no-ghost theorem

We prove a slight extension of the no-ghost theorem. The idea of using the no-ghost theorem to prove results about Kac-Moody algebras appears in Frenkel's paper [15], which also contains a proof of the no-ghost theorem. The original proof of Goddard and Thorn [21] works for the cases we need with only trivial modifications. For convenience we give a quick sketch of their proof.

Theorem 5.1. *(The no-ghost theorem.) Suppose that V is a vector space with a nonsingular bilinear form (,) and suppose that V is acted on by the Virasoro algebra in such a way that the adjoint of L_i is L_{-i}, the central element of the Virasoro algebra acts as multiplication by 24, any vector of V is a sum of eigenvectors of L_0 with nonnegative integral eigenvalues, and all the eigenspaces of L_0 are finite dimensional. We let V^i be the subspace of V on which L_0 has eigenvalue i. Assume that V is acted on by a group G which preserves all this structure. We let $V_{II_{1,1}}$ be the vertex algebra of the double cover $\hat{II}_{1,1}$ of the two dimensional even unimodular Lorentzian lattice $II_{1,1}$ (so that $V_{II_{1,1}}$ is $II_{1,1}$-graded, has a bilinear form (,), and is acted on by the Virasoro algebra as in section 3). We let P^1 be the subspace of the vertex algebra $V \otimes V_{II_{1,1}}$ of vectors v with $L_0(v) = v, L_i(v) = 0$ for $i > 0$, and we let P^1_r be the subspace of P^1 of degree $r \in II_{1,1}$. All these spaces inherit an action of G from the action of G on V and the trivial action of G on $V_{II_{1,1}}$ and $\mathbf{R}^2$. Then the quotient of P^1_r by the nullspace of its bilinear form is naturally isomorphic, as a G module with an invariant bilinear form, to $V^{1-(r,r)/2}$ if $r \neq 0$ and to $V^1 \oplus \mathbf{R}^2$ if $r = 0$.*

The name "no-ghost theorem" comes from the fact that in the original statement of the theorem in [21], V was part of the underlying vector space of the vertex algebra of a positive definite lattice so the inner product on V^i was positive definite, and thus P^1_r had no ghosts (i.e. vectors of negative norm) for r nonzero. The space V^i is the space V_{i-1} used in the rest of this paper.

Sketch of proof (taken from Goddard and Thorn, [21]). Fix some nonzero $r \in II_{1,1}$ and some norm 0 vector w in $II_{1,1}$ with $(r,w) \neq 0$.

We use the following operators. We have an action of the Virasoro algebra on $V \otimes V_{II_{1,1}}$ generated by its conformal vector. The operators L_i of the Virasoro algebra satisfy the relations

$$[L_i, L_j] = (i-j)L_{i+j} + 26(i^3 - i)\delta^i_{-j}/12.$$

and the adjoint of L_i is L_{-i}. (The 26 comes from the 24 in theorem 5.1 plus the dimension of $II_{1,1}$.) We define operators K_i, $i \in \mathbf{Z}$ by $K_i = v_{i-1}$ where v is the element $e^{-w}_{-2}e^w$ of the vertex algebra of $II_{1,1}$, and e^w is an element of the group ring of the double cover of $II_{1,1}$ corresponding to $w \in II_{1,1}$ and e^{-w} is its inverse. (There are in fact two possible choices for e^w which differ by a factor of -1, but it does not matter which we choose because this factor of -1 cancels out in the expression $e^{-w}_{-2}e^w$.) These operators satisfy the relations

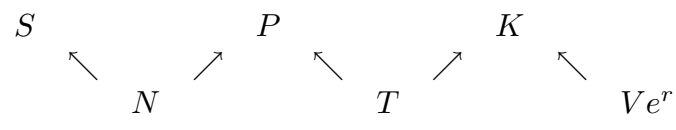
$$[L_i, K_j] = -jK_{i+j}$$
$$[K_i, K_j] = 0$$

because w has norm 0, and the adjoint of K_i is K_{-i}.

We define the following subspaces of $V \otimes V_{II_{1,1}}$.

- H is the subspace of $V \otimes V_{II_{1,1}}$ of degree $r \in II_{1,1}$. H^1 is its subspace of vectors h with $L_0(h) = h$.
- P, is the subspace of H of all vectors h with $L_i(h) = 0$ for $i > 0$. $P^1 = H^1 \cap P$.
- S, the space of spurious vectors, is the subspace of H of vectors perpendicular to P. $S^1 = H^1 \cap S$.
- N $= S \cap P$ is the radical of the bilinear form of P, and $N^1 = H^1 \cap N$.
- T, the transverse space, is the subspace of P annihilated by all the operators $K_i, i > 0$, and $T^1 = H^1 \cap T$.
- K is the space generated by the action of the operators $K_i, i \in \mathbf{Z}$ on T.
- Ve^r is the subspace $V \otimes e^r$ of H.

We have the following inclusions of subspaces of H:



and we construct the isomorphism from $V^{1-(r,r)/2}$ to $P^1/N \cap P^1$ by zigzagging up and down this diagram; more precisely we show that Ve^r and T are both isomorphic to K modulo its nullspace, and then we show that T^1 is isomorphic to P^1 modulo its nullspace $P^1 \cap N$.

Lemma 5.1. *If f is a vector of nonzero norm in T then the vectors of the form*

$$L_{m_1}L_{m_2}\ldots K_{n_1}K_{n_2}\ldots(f)$$

for all sequences of integers with $0 > m_1 \geq m_2 \ldots$, $0 > n_1 \geq n_2 \ldots$ are linearly independent and span a space invariant under the operators K_i and L_i on which the bilinear form is nonsingular.

Sketch of proof: It is possible to order these elements so that the matrix of inner products is upper triangular with nonzero diagonal elements. This implies that they are independent and span a space on which the inner product is nonsingular. The other statements in the lemma are easy to prove.

Lemma 5.2. *The bilinear form on T is nonsingular, and K is the direct sum of T and the nullspace of K.*

Proof. If we choose an orthogonal basis for T, then the set of all vectors generated from them as in Lemma 5.1 is a basis for H. As H is nonsingular, T must therefore also be nonsingular. It now follows that K is the direct sum of its nullspace and T, because K is generated from T by the operators K_i and anything generated from T by these operators is in the nullspace of K.

Lemma 5.3. *Ve^r is naturally isomorphic to T.*

Proof. K is also the subspace of H annihilated by all the operators K_i, $i > 0$ from which it follows that K is the direct sum of Ve^r and the nullspace of K. By lemma 5.2 K is also the direct sum of T and the nullspace of K, so this gives a natural isomorphism from Ve^r to T, because they are both isomorphic to the quotient of K by its nullspace.

Lemma 5.4. *The associative algebra generated by the elements L_i for $i < 0$ is generated by elements mapping S^1 into S.*

Proof. Calculation shows that the operators L_{-1} and $L_{-2}+3L_{-1}^2/2$ have this property, and they generate the algebra generated by all L_i's with $i < 0$. (It is in doing these calculations that we need to use the fact that the center of the Virasoro algebra acts as 24 on V and therefore as 26 on H.)

Lemma 5.5. *P^1 is the direct sum of T^1 and N^1.*

Proof. We have to show that any vector p of P^1 can be written as $p = t+n$ for $t \in T^1$, $n \in N^1$. By lemma 5.3, $p = k + s$ for some unique $k \in K^1$, $s \in S^1$ and by lemma 5.4 k and s are both mapped into K and S by some set of generators of the algebra generated by the L_i's, $i > 0$. As these generators annihilate p, they must also annihilate k and s. Hence k is in $K^1 \cap P^1 = T^1$ and s is in $S^1 \cap P^1 = N^1$, so that $P^1 = T^1 \oplus N^1$.

The no ghost theorem now follows immediately, because by lemma 5.3 $V^{1-(r,r)/2}e^r$ is isomorphic to T^1, and by lemma 5.5 T^1 is isomorphic to the quotient of P^1 by its nullspace.

6 Construction of the Monster Lie algebra.

We construct the monster Lie algebra M in the proof of the following theorem. This is a generalized Kac-Moody algebra acted on by the monster simple group and graded by $II_{1,1} = \mathbf{Z}^2$. Recall that V is the vertex algebra of the monster, so that V is graded with pieces of dimensions 1, 0, 196884, ... , and $V_{II_{1,1}}$ is the vertex algebra of the two dimensional even Lorentzian lattice $II_{1,1}$ with matrix $\left(\begin{smallmatrix} 0 & -1 \\ -1 & 0 \end{smallmatrix}\right)$.

Theorem 6.1. *There exists a Lie algebra M, which we will call the monster Lie algebra, with the following properties.*

(1) M is $II_{1,1}$-graded.

(2) M has a contravariant bilinear form $(,)_0$ which is positive definite on the piece of degree $(m,n) \neq (0,0)$. (By "contravariant bilinear form" we mean that the $II_{1,1}$-graded Lie algebra M has an involution ω which acts as -1 on $II_{1,1}$ and as -1 on

the piece of degree (0,0), such that the form $(u,v) = -(u,\omega(v))_0$ *is invariant, and* $(u,v) = 0$ *unless* $\deg(u) + \deg(v) = 0$.*)*

(3) M *is acted on by the monster. As a representation of the monster, the piece of* M *of degree* (m,n) *is isomorphic to* V_{mn} *if* $(m,n) \neq (0,0)$*, and to the trivial representation* $\mathbf{R}^2$ *if* $(m,n) = (0,0)$ *where* V_n *is the piece of the monster vertex algebra* V *of conformal weight* $n+1$.

Proof. The tensor product $V \otimes V_{II_{1,1}}$ of the vertex algebras V and $V_{II_{1,1}}$ is also a vertex algebra. If P^i is the space of vectors v of this vertex algebra satisfying $L_j(v) = 0$ if $j > 0$, $L_0(v) = v$, then P^1/DP^0 is a Lie algebra with an invariant bilinear form (,), and an involution ω induced by the trivial automorphism of V and the involution ω of $V_{II_{1,1}}$ defined in section 3. We define the monster Lie algebra M to be the quotient of the Lie algebra P^1/DP^0 by the kernel of the form (,). (The kernel of the bilinear form on P^1 is strictly larger than DP^0.)

The $II_{1,1}$ grading of the vertex algebra $V_{II_{1,1}}$ induces a $II_{1,1}$ grading on the Lie algebra M. The no-ghost theorem 5.1 implies that the piece of M of degree $(m,n) \in II_{1,1}$ is isomorphic to the piece of V of degree $1 - (m,n)^2/2 = 1 - mn$ if $v \neq 0$ and to $\mathbf{R}^2$ if $v = 0$, and that $(g,\omega(g)) > 0$ if $g \in M$ is nonzero and homogeneous of nonzero degree in $II_{1,1}$. This proves theorem 6.1.

Remark. The construction of a Lie algebra from a vertex algebra in theorem 6.1 can be carried out for any vertex algebra V with a conformal vector, but it is only when this vector has dimension 24 that we can apply the no-ghost theorem to identify the homogeneous pieces of M with those of V. The important point is that the bilinear form on M is positive definite on any piece of nonzero degree, and this need not be true if the conformal vector has dimension greater than 24, even if the inner product on V is positive definite.

Theorem 6.2. *The monster Lie algebra* M *is a generalized Kac-Moody algebra.*

Proof: This follows from theorem 6.1. The only condition in the definition of a generalized Kac-Moody algebra that is not immediately obvious from the conclusion of theorem 6.1 is the one about the **Z**-grading of M. We can **Z**-grade M in a suitable way by letting elements of M of degree $(m,n) \in II_{1,1}$ have degree $2m + n \in \mathbf{Z}$.

Remark. If V_L is the vertex algebra of any positive definite even lattice L of dimension at most 24 then the bilinear form on the corresponding Lie algebra M_L is still almost positive definite, so M_L is a generalized Kac-Moody algebra. If V_Λ is the vertex algebra of the Leech lattice Λ then M_Λ is (by definition) the fake monster Lie algebra, and its simple roots are described explicitly in [8] (where it is called the monster Lie algebra, or more precisely proved to be isomorphic to something called the monster Lie algebra there). Section 12 contains a summary of results about the fake monster Lie algebra. I have not found any other lattices for which the simple roots of the corresponding Lie algebra can be found explicitly, although for the lattice $E_8 \oplus E_8$ all the simple roots of norm 0, -2, or -4 have multiplicity 0 or 1 so there may be something interesting going on in this case. There are some calculations for the case when V_L is the vertex algebra of $L = E_8$ in Kac, Moody and Wakimoto [24].

The monster Lie algebra can also be constructed as a semi-infinite cohomology group; see [19].

The construction of the monster Lie algebra above looks bizarre at first sight, so we briefly explain some of the motivation behind it. The fake monster vertex algebra M_Λ [8] is the Lie algebra P^1/(kernel of bilinear form) of the vertex algebra of the lattice $II_{25,1}$. This lattice is the sum of the Leech lattice Λ and the lattice $II_{1,1}$, and the vertex algebra of the sum of two lattices Λ and $II_{1,1}$is the tensor product of the vertex algebras V_Λ and $V_{II_{1,1}}$ of the lattices, so the fake monster Lie algebra is the space P^1/kernel of $V_\Lambda \otimes V_{II_{1,1}}$. The monster vertex algebra V is similar to the vertex algebra V_Λ, which suggests replacing V_Λ by V in the construction above, and this gives the monster Lie algebra M.

7 The simple roots of the monster Lie algebra.

We find the simple roots of the monster Lie algebra M constructed in section 6. This algebra turns out to have a Weyl vector, i.e., a vector ρ with $(\rho, r) = -(r, r)/2$ for all simple roots r.

We find the simple roots of the monster Lie algebra by using the following identity, which turns out to be its denominator formula.

Lemma 7.1.

$$p^{-1} \prod_{\substack{m>0 \\ n\in\mathbf{Z}}} (1 - p^m q^n)^{c(mn)} = j(p) - j(q) \tag{7.1}$$

where $j(q) - 744 = \sum_n c(n)q^n = q^{-1} + 196884q + \ldots$.

Proof. This is essentially just lemma 2 of section 4 of [8]. For convenience we recall the proof. If we multiply the left hand side by p and take its logarithm we get

$$\begin{aligned}
&- \sum_{m>0} \sum_{n\in\mathbf{Z}} \sum_{k>0} c(mn) p^{mk} q^{nk}/k \\
&= - \sum_{m>0} \sum_{n\in\mathbf{Z}} \sum_{0<k|(m,n)} \frac{1}{k} c(\frac{mn}{k^2}) p^m q^n \\
&= - \sum_{m>0} T_m (\sum_{n\in\mathbf{Z}} c(n) q^n) p^m \\
&= - \sum_{m>0} T_m (j(q) - 744) p^m \\
&= \sum_{m>0} f_m(q) p^m
\end{aligned}$$

where T_m is the m'th Hecke operator (see Serre [30] proposition 12, chapter VII, section 5) and each f_m is a modular function of level 1, holomorphic on the upper half plane, and therefore a polynomial in $j(q)$. If we exponentiate this we see that the left hand side of 7.1 is of the form $\sum_{m\geq -1} g_m(q)p^m$ where each g_m is a polynomial in $j(q)$. Each coefficient of p^m of the right hand side of (7.1) is either a constant (if $m \neq 0$) or $744 - j(q)$ (if $m = 0$) and is therefore also a polynomial in $j(q)$. Any polynomial in $j(q)$ is determined by its coefficients of q^n for $n \leq 0$, so to prove the lemma it is sufficient to check that the coefficients of $q^n, n \leq 0$ of both sides are the same, which is easy to do. This proves lemma 7.1.

Remark. The infinite product in (7.1) converges if $|p|, |q| < e^{-2\pi}$ and $p \neq q$, and the infinite series for $j(q)$ converges if $|q| < 1$.

Theorem 7.2. *The simple roots of the monster Lie algebra M are the vectors $(1, n)$ ($n = -1$ or $n > 0$), each with multiplicity $c(n)$.*

Let N be the generalized Kac-Moody algebra with root lattice $II_{1,1}$ whose simple roots are as stated in the theorem. The simple roots of a generalized Kac-Moody algebra (with given Cartan subalgebra and choice of fundamental Weyl chamber) are determined by its root multiplicities because of the denominator formula, so it is sufficient to show that the multiplicity of the root (m, n) of N is equal to the multiplicity $c(mn)$ of the root (m, n) of the algebra M (so N is isomorphic to M). The denominator formula of the Lie algebra N is

$$e^\rho \prod_{\substack{m>0 \\ n\in\mathbf{Z}}} (1 - p^m q^n)^{\mathrm{mult}(m,n)} = \sum_{w\in W} \det(w) w(e^\rho S) \tag{7.2}$$

where $\mathrm{mult}(m, n)$ is the multiplicity of the root (m, n) of N, p and q are the elements $e^{(1,0)}$ and $e^{(0,1)}$ of the group ring of $II_{1,1}$, W is the Weyl group of N which is of order 2 and whose nontrivial element exchanges p and q, and e^ρ is $e^{(-1,0)} = p^{-1}$. By the denominator formula quoted in section 4 the element S is

$$S = \sum_A (-1)^{|A|} e^{\Sigma A}$$

where the sum is over all finite subsets A of the set of simple imaginary roots such that any two distinct elements of A are orthogonal, $|A|$ is the number of elements of A, and ΣA is the sum of the elements of A. All the imaginary simple roots of N have nonzero inner product with each other, so S is equal to $1 - \sum_{n>0} c(n) p q^n = 1 - p(j(q) - q^{-1} - 744)$ and hence the right hand side of the denominator formula (7.2) of N is

$$\begin{aligned} e^\rho S - w(e^\rho S) &= (p^{-1} - j(q) + q^{-1} + 744) - (q^{-1} - j(p) + p^{-1} + 744) \\ &= j(p) - j(q) \end{aligned}$$

where w is the nontrivial element of the Weyl group W. This is the right hand side of (7.1), so that the left hand sides of (7.1) and (7.2) must also be the same, which implies that the multiplicity of the root $(m, n) \neq (0, 0)$ of N is $c(mn)$. This proves theorem 7.2.

Remark. This shows that the action of the monster on M can be extended to an action of the full automorphism group of the graded vector space V on M which still preserves the grading and Lie algebra structure of M. The special property of the action of the monster on M that we use is that the pieces of degree (a, b) and (c, d) of the $II_{1,1}$-graded Lie algebra M are isomorphic as representations of the monster whenever $ab = cd \neq 0$.

8 The twisted denominator formula.

The monster Lie algebra, like any generalized Kac-Moody algebra, can be written as a direct sum $E \oplus H \oplus F$ of subalgebras, where H is the Cartan subalgebra and E and

F are the subalgebras corresponding to the positive and negative roots. We recall from section 4 the equality

$$\Lambda(E) = H(E) \tag{8.1}$$

where in the case of the monster Lie algebra both sides are virtual $II_{1,1}$-graded modules over the monster. In this section we calculate both sides of (8.1) explicitly and use this to obtain some relations between the coefficients of the Thompson series $T_g(q)$.

For any generalized Kac-Moody algebra the homology groups can be calculated as in section 4 provided we know the simple roots. In the case of the monster Lie algebra M the simple roots are given in section 7 and are just the vectors $(1, n)$, $n = -1$ or $n > 0$ with multiplicity $c(n)$, so we can calculate the homology groups $H_i(E)$ as follows. Recall that $H_i(E)$ is the subspace of elements of $\Lambda^i(E)$ whose degree r (in the root lattice) satisfies $(r, r + 2\rho) = 0$. For the monster Lie algebra we can think of the roots as elements of $II_{1,1}$ and can identify the Weyl vector ρ with the norm 0 vector $(-1, 0)$. The condition on r then says that $r + \rho$ has norm 0, i.e., r is a vector of the form $(m, 0)$ or $(1, n)$, so we just have to find the elements of $\Lambda^i(E)$ of these degrees. The answers we get are as follows (where we write $M_{m,n}$ for the subspace of elements of M of degree $(m, n) \in II_{1,1}$).

$H_0(E)$ is one dimensional with character 1, because $\Lambda^0(E) = \mathbf{R}$ is just a one dimensional space of degree $(0, 0)$.

$H_1(E)$ is the subspace of $\Lambda^1(E) = E$ of elements of degree $(m, 0)$ or $(1, n)$. There are no elements of degree $(m, 0)$ and the space of elements of degree $(1, n)$ is the simple root space $M_{1,n} \cong V_n$, so $H_1(E)$ has character $p(j(q) - 744)$ and its piece of degree $(1, n)$ is isomorphic to the n'th head representation V_n of the monster. (Notice that the first homology group $H_1(E)$ is isomorphic to the sum of the simple root spaces; the same is true for any generalized Kac-Moody algebra.)

For $i \geq 2$ there are no elements in $\Lambda^i(E)$ of degree $(1, n)$ because all elements of E have degrees of the form (m, n) with $m \geq 1$. Moreover they all have degrees with $n > 0$ except for the one dimensional space with degree equal to the real simple root $(1, -1)$, so the only way for an element of $\Lambda^i(E)$ to have degree $r = (m, 0)$ is for it to be the exterior product of two elements of E of degrees $(1, -1)$ and $(m - 1, 1)$. Therefore $H_2(E)$ is the sum of pieces of degrees $(m, 0)$ for $m \geq 2$ isomorphic to $M_{m-1,1} = V_{m-1}$, and $H_i(E) = 0$ for $i \geq 3$.

To summarize, the homology groups of E are the following $II_{1,1}$-graded vector spaces, where we use p and q to stand for one dimensional vector spaces of degrees $(1, 0)$ and $(0, 1)$ in $II_{1,1}$, and V_i has degree $0 \in II_{1,1}$ for all i.

$H_0(E) = V_{-1} = \mathbf{R}$ and has character 1.
$H_1(E) = \sum_{n \in \mathbf{Z}} V_n pq^n$ and has character $\sum_{n \in Z} c(n) pq^n = p(j(q) - 744)$.
$H_2(E) = p \sum_{m>0} V_m p^m$ and has character $p(j(p) - 744) - 1$.
$H_i(E) = 0$ if $i \geq 3$.

For most interesting generalized Kac-Moody algebras there are an infinite number of nonzero homology groups. In general the largest n with H_n nonzero is equal to (maximum length of an element of the Weyl group) + (maximum number of pairwise orthogonal imaginary simple roots) which is usually infinite; however for the monster Lie algebra both the terms in parentheses happen to be 1.

The alternating sum $\oplus(-1)^i H_i(E)$ is therefore equal to $p(\sum_m V_m p^m - \sum_n V_n q^n)$, where we use p and q to stand for one dimensional vector spaces of degrees (1,0) and (0,1). If we substitute these values of the homology groups into (8.1) we find that

$$p^{-1}\Lambda(\sum_{m>0,n\in\mathbf{Z}} V_{mn}p^m q^n) = \sum_m V_m p^m - \sum_n V_n q^n \tag{8.2}$$

where V_m is the m'th head representation of the monster simple group. (Both sides of (8.2) are essentially $II_{1,1}$-graded virtual representations of the monster simple group.) For any finite dimensional vector space U, $\Lambda(U)$ is naturally isomorphic to $\exp(-\sum_{i>0}\psi^i(U)/i)$ (where ψ^i is the i'th Adams operation) by the splitting principle, because both expressions are multiplicative in U and are equal if U is one dimensional. (The splitting principle says that two natural operations on representations are equal if they are equal on sums of 1-dimensional representations, i.e., we can pretend that any representation is a sum of 1-dimensional ones, and the Adams operation ψ^i on virtual representations of a group G is defined by $\mathrm{Tr}(g|\psi^i(U)) = \mathrm{Tr}(g^i|U)$; see [2] for more details.) The same is true for infinite dimensional vector spaces U graded by some lattice L provided that the homogeneous pieces are all finite dimensional, and the pieces of degree $a \in L$ vanish unless a is in some fixed closed cone which does not contain any line and has its vertex at the origin; this condition ensures that all the virtual vector spaces $\Lambda(U)$, $\psi^i(U)$, and so on are graded with finite dimensional pieces of each degree and are therefore meaningful. Therefore if we take the trace of some element g of the monster on both sides of (8.2) we find that

$$p^{-1}\exp(-\sum_{i>0}\sum_{m>0,n\in\mathbf{Z}} \mathrm{Tr}(g^i|V_{mn})p^{mi}q^{ni}/i) = \sum_{m\in\mathbf{Z}} \mathrm{Tr}(g|V_m)p^m - \sum_{n\in\mathbf{Z}} \mathrm{Tr}(g|V_n)q^n \tag{8.3}$$

where $\mathrm{Tr}(g|V_n)$ is the trace of g on the vector space V_n.

These relations between the coefficients of the functions $\sum \mathrm{Tr}(g|V_n)q^n$ are the same as those that Norton [28] conjectured to hold for the modular functions associated with elements of the monster in Conway and Norton [13, table 2]. The identities 8.3 imply that the functions $T_g(q)$ are completely replicable in the terminology of [28]; in fact the definition of a completely replicable function in [28] is that the function should satisfy some identities equivalent to 8.3. Norton's conjectures that these modular functions are completely replicable were proved by Koike [25]. In the next section we use this to prove that the functions $\sum_m \mathrm{Tr}(g|V_m)q^m$ are these modular functions.

9 The moonshine conjectures.

In this section we complete the proof of theorem 1.1, i.e., we verify that the monster vertex algebra V satisfies Conway and Norton's moonshine conjectures [13] that the Thompson series of elements of the monster are certain Hauptmoduls. To do this we show that the identities (8.3) in section 8 imply that the Thompson series are determined by their first 5 coefficients $c_g(i)$, $1 \le i \le 5$, and then use the fact that the Hauptmoduls listed by Conway and Norton in [13, table 2] satisfy the same identities ([25]) and have the same first five coefficients.

The coefficients $c_g(n) = \mathrm{Tr}(g|V_n)$ of the Thompson series of elements of the monster satisfy the relations (8.3). If we compare the coefficients of p^2 and p^4 of both sides of 8.3 and carry out some elementary algebra we find that the coefficients $c_g(i)$ satisfy the recursion formulas below for $k \ge 1$:

$$\begin{aligned}
c_g(4k) &= c_g(2k+1) + (c_g(k)^2 - c_{g^2}(k))/2 + \sum_{1\le j<k} c_g(j)c_g(2k-j)\\
c_g(4k+1) &= c_g(2k+3) - c_g(2)c_g(2k) + (c_g(2k)^2 + c_{g^2}(2k))/2\\
&\quad + (c_g(k+1)^2 - c_{g^2}(k+1))/2 + \sum_{1\le j\le k} c_g(j)c_g(2k-j+2)\\
&\quad + \sum_{1\le j<k} c_{g^2}(j)c_g(4k-4j) + \sum_{1\le j<2k} (-1)^j c_g(j)c_g(4k-j)\\
c_g(4k+2) &= c_g(2k+2) + \sum_{1\le j\le k} c_g(j)c_g(2k-j+1)\\
c_g(4k+3) &= c_g(2k+4) - c_g(2)c_g(2k+1)\\
&\quad - (c_g(2k+1)^2 - c_{g^2}(2k+1))/2 + \sum_{1\le j\le k+1} c_g(j)c_g(2k-j+3)\\
&\quad + \sum_{1\le j\le k} c_{g^2}(j)c_g(4k-4j+2) + \sum_{1\le j\le 2k} (-1)^j c_g(j)c_g(4k-j+2)
\end{aligned} \tag{9.1}$$

where $c_g(n) = \mathrm{Tr}(g|V_n)$, $c_{g^2}(n) = \mathrm{Tr}(g^2|V_n)$. In particular if $n = 4$ or $n > 5$ then the coefficient $c_g(n)$ is determined by the coefficients $c_g(i)$ and $c_{g^2}(i)$ for $1 \le i < n$, so if we know all the coefficients $c_g(n)$ for $n = 1$, 2, 3, and 5 and all elements g of the monster then we can work out all the coefficients $c_g(n)$. (The coefficient $c_g(5)$ is not determined by the recursion relations because they degenerate into $c_g(5) = c_g(5)$. The Thompson series of an element of the monster is not determined by its first 5 coefficients $c_g(i)$, $1 \le i \le 5$, but is determined by the coefficients $c_{g^j}(i)$, $1 \le i \le 5$ of the Thompson series of all its powers g^j. For example, the Thompson series of the elements $60F$ and $93A$ (in ATLAS [14] notation) have the same first five coefficients $c_g(i)$, $1 \le i \le 5$. Norton showed in [28] that any completely replicable function $\sum a(n)q^n$ is determined by its 12 coefficients a_i, $i = 1$, 2, 3, 4, 5, 7, 8, 9, 11, 17, 19, 23.)

For the case of the elliptic modular function $j(q)$, where $g = 1$, these recursion formulas were discovered by Mahler [27]. We do not make any use of the formulas (9.1), apart from the fact that the coefficients for $n = 1$, 2, 3 and 5 determine all the coefficients. (This

is easy to check without calculating the formulas (9.1) explicitly. The reason for listing them explicitly is that they are a very efficient way of working out the coefficients c_g on a computer.)

Norton [28] stated and Koike [25] proved that the modular functions associated with elements of the monster in [13, table 2] also satisfy the relations (8.3) (with $\mathrm{Tr}(g|V_m)$ replaced by the coefficient of q^m of the modular function corresponding to g) and therefore also satisfy the recursion relations (9.1) above. To prove that the Thompson series of elements of the monster are these modular functions, it is therefore sufficient to check that the coefficients of q^i, $i \le 5$ of each function are the same.

The coefficients of q^i, $i \le 5$ of each modular function are in Conway and Norton [13, table 4], and they define representations of the monster which decompose as stated at the end of section 3. We can work out the coefficients $c_g(i)$, $1 \le i \le 5$ of the Thompson series of elements of the monster as follows. Knowing the coefficients $c_g(i)$, $1 \le i \le 5$ of the Thompson series is equivalent to knowing how the representations $V_i, 1 \le i \le 5$ of the monster decompose into irreducible representations of the monster. The only irreducible representations of the monster with dimension at most that of V_5 are the first seven, with characters χ_i, $1 \le i \le 7$ in atlas [14] notation. Therefore we can evaluate the coefficients $c_g(i)$, $1 \le i \le 5$ of the Thompson series of all elements of the monster provided we can find seven elements g_j, $1 \le j \le 7$ of the monster for which we can evaluate these coefficients, and such that the matrix $\chi_i(g_j)$, $1 \le i, j \le 7$ is nonsingular, because this determines the decomposition of V_i, $1 \le i \le 5$, into irreducible representations of the monster.

In their book [16] Frenkel, Lepowsky and Meurman give an explicit formula for the Thompson series of any element of the centralizer of an involution of type 2B of the monster. The elements of odd order of $2^{1+24}.Co_1$ correspond to the elements of odd order of the group $2.Co_1$ of automorphisms of the Leech lattice Λ. Frenkel, Lepowsky and Meurman's formula is particularly easy to evaluate if the corresponding automorphism of the Leech lattice is of odd order and fixes no nonzero vectors: it becomes

$$2\sum_n \mathrm{Tr}(g|V_n)q^n = 1/\eta_g(q) + \eta_g(q)/\eta_g(q^2) + \eta_g(q)/\eta_g(q^{1/2}) + \eta_g(q)/\eta_g(-q^{1/2}) \qquad (9.2)$$

where $\eta_g(q)$ is the eta function of g acting on the Leech lattice, equal to $\eta(\epsilon_1 q)\dots\eta(\epsilon_{24}q)$ if g has eigenvalues $\epsilon_1, \dots, \epsilon_{24}$ on the vector space $\Lambda \otimes \mathbf{R}$, where $\eta(q) = q^{1/24}\prod_{n>0}(1-q^n)$. If we put $f(\tau) = 1/\eta_g(q)$ (where $q = e^{2\pi i\tau}$) then $f(\tau) - f(0)$ is the modular function of genus 0 that Conway and Norton associate to the element g of the monster in [13, table 2]. The coefficients of the right hand side of the formula above are easy to evaluate explicitly, and we can check that its coefficients of q^i for $i \le 5$ are equal to those of the modular function $f(\tau) - f(0)$. (The right hand side of 9.2 and $f(\tau) - f(0)$ are of course equal, as we can see directly by observing that the right hand side of (9.2) is $f(\tau) + T_2 f(\tau)/f(\tau)$, where T_2 is a Hecke operator.)

For our seven elements g_j of the monster we choose an element g_1 of conjugacy classes 2B (for which Frenkel, Lepowsky and Meurman explicitly evaluate the Thompson series), and 6 more elements g_i of the monster corresponding to odd order automorphisms of the Leech lattice with no nonzero fixed vectors such that the determinant of the 7×7 matrix $\chi_i(g_j)$ is nonzero. A set of 6 elements satisfying these conditions are those of type

3B, 5B, 7B, 9B, 13B, 15D in the monster (using atlas [14] notation) corresponding to elements in the conjugacy classes 3A, 5A, 7A, 9A, 13A, and 15C of Co_1 with cycle shapes $3^{12}1^{-12}$, 5^61^{-6}, 7^41^{-4}, 9^31^{-3}, 13^21^{-2}, and 15^23^{-2}. The determinant of the matrix $\chi_i(g_j)$ is $35672555520 = 2^{22}3^55^17^1$ which is nonzero. The modular groups corresponding to these elements of the monster are $\Gamma_0(3)$, $\Gamma_0(5)$, $\Gamma_0(7)$, $\Gamma_0(9)$, $\Gamma_0(13)$, and $\binom{3,0}{0,1}^{-1}\Gamma_0(5)\binom{3,0}{0,1}$. (It is necessary to use at least one element g_i of even order, because the characters of the first seven representations of the monster are linearly dependent when restricted to elements of odd order; in fact the character of $S^2(V_1) - \Lambda^2(V_1) - V_1$ is 0 on all elements of odd order.)

We can use this to verify that the representations V_1, V_2, V_3, and V_5 of the monster decompose as stated at the end of section 3, and this implies that the numbers $\mathrm{Tr}(g|V_n)$ are equal to the coefficients of the corresponding modular functions in [13] for $n = 1$, 2, 3 and 5, and therefore for all n by the recursion relations (9.1). This completes the verification that the Thompson functions $\sum_n \mathrm{Tr}(g|V_n)q^n$ are modular functions of genus 0 and proves theorem 1.1.

10 The monstrous Lie superalgebras.

In the remainder of the paper we describe some Lie algebras similar to the monster Lie algebra. In this section we describe some associated to elements of the monster.

The twisted denominator formula for some diagram automorphism g of a generalized Kac-Moody algebra, which is obtained by taking the trace of g on both sides of the identity $H(E) = \Lambda(E)$, is often the untwisted denominator formula for some generalized Kac-Moody algebra or superalgebra. For example, the twisted denominator formula for E_6 is just the ordinary denominator formula for F_4. We construct a family of Lie algebras and superalgebras whose denominator formulas are twisted denominator formulas of the monster Lie algebra. (Warning: unlike the case of finite dimensional Lie algebras, these are not always subalgebras of the monster Lie algebra; in fact some of them are Lie superalgebras.)

If g is an element (or conjugacy class) of the monster group M then we write $T_g(q)$ for the Thompson series $T_g(q) = \sum c_g(n)q^n = \sum \mathrm{Tr}(g|V_n)q^n$ of g. For example, $T_1(q) + 744$ is the elliptic modular function $j(q)$. The function $T_g(q)$ is the normalized generator for a genus zero function field of a group containing $\Gamma_0(nh)$, where n is the order of g and h is an integer with $h|(24, n)$.

We use the convention that for super vector spaces the dimension is defined to be the dimension of the even part minus the dimension of the odd part. For example, the multiplicity of a root is the dimension of the root space, and is therefore equal to the dimension of the even root space minus the dimension of the odd root space. This provides an interpretation of roots of negative multiplicity. With this convention for multiplicities the denominator formula for generalized Kac-Moody superalgebras is the same as for generalized Kac-Moody algebras.

We define the generalized Kac-Moody superalgebra of an element g of the monster to have root lattice $II_{1,1}$ and simple roots $(1, n)$ with multiplicity $\mathrm{Tr}(g|V_n)$. (These multiplicities are all integers as can be seen in several ways; for example they are rational by the recursion relations 9.1 and algebraic integers because they are character values.) This is really a family of Lie superalgebras rather than a single superalgebra, because for each

simple root $(1,n)$ we can choose any two nonnegative integers a and b with $a-b = \mathrm{Tr}(g|V_n)$ and take a even simple roots $(1,n)$ and b odd ones. This only seems to be interesting if all coefficients of T_g are positive, in which case we get a Lie algebra, or if the coefficients alternate in sign, in which case we get a superalgebra such that the super elements are those of degree (m,n) with $m+n$ odd.

We can work out the root multiplicities of these superalgebras because their denominator formulas are the twisted denominator formulas of the monster Lie algebra. The denominator formula for the Lie superalgebra of g is

$$p^{-1}\prod_{m>0,n\in\mathbf{Z}}(1-p^mq^n)^{\mathrm{mult}(m,n)} = \sum_m \mathrm{Tr}(g|V_m)p^m - \sum_n \mathrm{Tr}(g|V_n)q^n = T_g(p)-T_g(q) \quad (10.1)$$

and by (8.3)

$$T_g(p) - T_g(q) = p^{-1}\exp(-\sum_{i>0}\sum_{m>0}\sum_{n\in\mathbf{Z}}\mathrm{Tr}(g^i|V_{mn})p^{mi}q^{ni}/i). \quad (10.2)$$

If we compare the logarithms of (10.1) and (10.2) we find that

$$\sum_{i>0,m>0,n\in\mathbf{Z}}\mathrm{Tr}(g^i|V_{mn})p^{mi}q^{ni}/i = \sum_{i>0,m>0,n\in\mathbf{Z}}\mathrm{mult}(m,n)p^{mi}q^{ni}/i \quad (10.3)$$

and applying the Moebius inversion formula to (10.3) shows that

$$\mathrm{mult}(m,n) = \sum_{ds|(m,n)}\frac{\mu(s)}{ds}\mathrm{Tr}(g^d|V_{mn/d^2s^2}) \quad (10.4)$$

where μ is the Moebius function. If f is any function of n whose value depends only on (n,N) for some N, then the sum $\sum_{ds=k}\mu(s)f(d)$ is 0 unless $k|N$. The trace $\mathrm{Tr}(g^d|V_n)$ depends only on n and (d,N) where N is the order of g, so $\sum_{ds=k}\mu(s)\mathrm{Tr}(g^d|V_{mn/d^2s^2})$ is 0 unless $k|N$, hence (10.4) can be simplified to

$$\mathrm{mult}(m,n) = \sum_{ds|(m,n,N)}\frac{\mu(s)}{ds}\mathrm{Tr}(g^d|V_{mn/d^2s^2}) \quad (10.5)$$

This is the formula for the multiplicity of the roots of the superalgebra of g. The sum on the right of (10.5) can often be simplified by expressing T_{g^d} in terms of T_g. For example, suppose that $N = N_1N_2$ is a squarefree integer such that the group $\Gamma_0(N)+d|N_1$ has genus 0, where $\Gamma_0(N)+d|N_1$ is the group generated by $\Gamma_0(N) = \{\left(\begin{smallmatrix}a & b\\ c & d\end{smallmatrix}\right) \in SL_2(\mathbf{Z})|c \equiv 0 \bmod N\}$ together with the Atkin-Lehner involutions of all divisors of N_1. (See [13].) If g is an element of the monster corresponding to one of these groups then so is g^k. By the compression formula of [13] section 8

$$T_{g^p}(\tau) = T_g(\tau) + T_g(\tau/p) + \ldots + T_g((\tau+p-1)/p)$$

for any prime divisor p of N_1, and

$$T_{g^p}(\tau) = T_g(\tau) - T_g(\tau/p^2) - \ldots - T_g((\tau + p^2 - 1)/p^2)$$

for any prime divisor p of N_2. These imply that

$$\mathrm{Tr}(g^d|V_n) = \sum_{d_1|(d,N_1)} \sum_{d_2|(d,N_2)} d_1 d_2^2 \mu(d_2) \mathrm{Tr}(g|V_{d_1 d_2^2 n}) \tag{10.6}$$

by induction on the number of divisors of d. If we substitute (10.6) into (10.5) and rearrange we find that

$$\mathrm{mult}(m,n) = \sum_{d|(m,n,N_1)} \mathrm{Tr}(g|V_{mn/d}) \prod_{p|(m,n,N_2)} (1-p) \tag{10.7}$$

where the product is over all primes p dividing (m, n, N_2).

We get particularly interesting results when N_2 is 1 or 2. First suppose that $N_2 = 1$ so that $N = N_1$ is one of the 44 squarefree integers that are orders of elements in the monster, which are the same as the integers for which $\Gamma_0(N)+$ has genus 0. Let g be an element of the monster corresponding to the group $\Gamma_0(N)+$, so that $T_g(q)$ is the normalized generator of the function field of $\Gamma_0(N)+$. Then (10.7) becomes

$$\mathrm{mult}(m,n) = \sum_{d|(m,n,N)} \mathrm{Tr}(g|V_{mn/d}) \tag{10.8}$$

and if we substitute (10.8) into (10.1) we obtain the formula

$$T_g(p) - T_g(q) = p^{-1} \prod_{m>0, n\in\mathbf{Z}} \prod_{d|N} (1 - p^{dm} q^{dn})^{c_g(dmn)}$$

which generalizes the product formula for the j function. For these cases the coefficients c_g of the Thompson series $T_g(q)$ are always nonnegative, so the algebras we get are Lie algebras rather than Lie superalgebras. This seems to happen whenever the corresponding subgroup of $PGL_2(R)$ contains the Fricke involution $\tau \to -1/N\tau$. More generally, Norton has observed ([29]) that the result of applying the Fricke involution to a Thompson series always has nonnegative coefficients.

Similarly suppose that $N_2 = 2$ so that N_1 is one of the squarefree integers 1, 3, 5, 7, 11, or 15 for which $\Gamma_0(2N_1) + d|N_1$ has genus 0. Then (10.7) becomes

$$\mathrm{mult}(m,n) = (-1)^{(m-1)(n-1)} \sum_{d|(m,n,N_1)} \mathrm{Tr}(g|V_{mn/d}).$$

The coefficients of the Thompson series for these six cases alternate in sign, so the multiplicity of (m, n) is at least 0 or at most 0 depending on whether $m + n$ is even or odd. Therefore the root (m, n) is an even root or an odd (super) root depending on the parity of $m + n$. The same is true for any element whose Thompson series has coefficients that alternate in sign. If the coefficients of a Thompson series are neither at least 0 nor alternating in sign, then the root space of some root (m, n) is in general a sum of an even space and an odd space, and the multiplicity of the root is the difference of the dimensions of these two spaces. However the Lie superalgebras are probably not very interesting unless the coefficients are either all positive or alternating in sign.

11 Some modular forms.

We verify some identities involving modular forms of level 2 that we use later. This section can be missed out if the reader is willing to assume these identities.

The modular forms in this section are usually forms of level 2, i.e. modular forms for the group $\Gamma(2) = \{\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in SL_2(\mathbf{Z}) | \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \equiv \left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right) \bmod 2\}$. We recall a few facts about this group from [22]. The group $\Gamma(2)$ has no elliptic elements, 3 cusps (represented by 0, 1, and ∞), genus 0, and has index 6 in $SL_2(\mathbf{Z})$. The dimension of the space of forms of even nonnegative weight $2k$ is equal to $k+1$ and if f and g are two linearly independent forms of weight 2 then every form can be written uniquely as a polynomial in f and g. Two modular forms of level 2 and weight $2k$ are equal if and only if their coefficients of q^0, $q^{1/2}$, q^1, ... , q^k are equal. The group $\Gamma(2)$ is conjugate to the group $\Gamma_0(4) = \{\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in SL_2(\mathbf{Z}) | c \equiv 0 \bmod 4\}$, and $f(q)$ is a modular form for $\Gamma_0(4)$ if and only if $f(q^{1/2})$ is a modular form for $\Gamma(2)$.

We list several modular forms of level 2, together with enough of their Fourier coefficients to verify the identities we use in later sections. Recall that to prove that two modular forms are equal it is sufficient to check that sufficiently many of their Fourier coefficients are equal.

The following are some forms of weight 4 and level 2.

$$\begin{aligned}
\theta_{E_8}(q) &= 1 + 240(q + 9q^2 + 28q^3 + \ldots + \sigma_3(n)q^n + \ldots) \\
\theta_{2E_8}(q) &= 1 + 240q^2 + \ldots \\
\theta_{2E_8+v_1}(q) &= 2q^{1/2} + 56q^{3/2} + 252q^{5/2} + \ldots \\
\theta_{2E_8+v_2}(q) &= 16q + 128q^2 + 448q^3 + \ldots \\
\eta(q)^{16}\eta(q^{1/2})^{-8} &= q^{1/2} + 8q + 28q^{3/2} + 64q^2 + 126q^{5/2} + 224q^3 + \ldots \\
\eta(q)^{16}\eta(-q^{1/2})^{-8} &= -q^{1/2} + 8q - \ldots \\
\eta(q)^{16}\eta(q^2)^{-8} &= 1 - 16q + 112q^2 - 448q^3 + \ldots
\end{aligned}$$

The form $\theta_{E_8}(q)$ is the theta function of the E_8 lattice, and the form $\theta_{2E_8+v_i}$ is the theta function of the coset of $2E_8$ containing a vector v_i of norm i in the dual of $2E_8$ but not in E_8 (where $2E_8$ is the E_8 lattice with all norms doubled).

Lemma 11. 1. *If* $\eta_{2+}(q) = \eta(q)^8\eta(q^2)^8$ *then*

$$\begin{aligned}
\theta_{2E_8}(q)/\Delta(q) &= 1/\eta_{2+}(q) + 1/\eta_{2+}(q^{1/2}) + 1/\eta_{2+}(-q^{1/2}) \\
\theta_{2E_8+v_1}(q)/\Delta(q) &= 1/\eta_{2+}(q^{1/2}) - 1/\eta_{2+}(-q^{1/2}) \\
\theta_{2E_8+v_2}(q)/\Delta(q) &= 1/\eta_{2+}(q^{1/2}) + 1/\eta_{2+}(-q^{1/2})
\end{aligned}$$

Proof: If we multiply both sides by $\Delta(q)$ then both sides are modular forms of level 2 and weight 4, so it is sufficient to check that the coefficients of q^0, $q^{1/2}$, and q^1 of both sides are equal, which we can do using the expressions above.

The following are some forms of weight 8 and level 2.

$$\begin{aligned}
\theta_{\Lambda_2}(q) &= 1 + 4320q^2 + \dots \\
\theta_{\Lambda_2}(q^{1/2}) &= 1 + 4320q^1 + 61440q^{3/2} + 522720q^2 + \dots \\
\theta_{\Lambda_2}(-q^{1/2}) &= 1 + 4320q - 61440q^{3/2} + 522720q^2 + \dots \\
\eta(q)^{32}\eta(q^2)^{-16} &= 1 - 32q + 480q^2 - 4480q^3 + \dots \\
\eta(q)^{32}\eta(q^{1/2})^{-16} &= q + 16q^{3/2} + 120q^2 + 576q^{5/2} + 2060q^3 + \dots \\
\eta(q)^{32}\eta(-q^{1/2})^{-16} &= q - 16q^{3/2} + 120q^2 \dots
\end{aligned}$$

The form $\theta_{\Lambda_2}(q)$ is the theta function of the Barnes-Wall lattice Λ_2 of dimension 16, and is a modular form for $\Gamma_0(2)+$. The forms $\theta_{\Lambda_2+v_i}$ are defined in the same way as for the lattice $2E_8$.

Lemma 11. 2. *Let $\eta_{2-}(q)$ be the modular form $\eta(q)^{-8}\eta(q^2)^{16}$. Then*

$$\begin{aligned}
\theta_{\Lambda_2}(q)/\Delta(q) &= 1/\eta_{2-}(q) - 1/\eta_{2-}(q^{1/4}) - 1/\eta_{2-}(iq^{1/4}) - 1/\eta_{2-}(-q^{1/4}) - 1/\eta_{2-}(-iq^{1/4}) \\
\theta_{\Lambda_2+v_3}/\Delta(q) &= 1/\eta_{2-}(q^{1/4}) + 1/\eta_{2-}(iq^{1/4}) + 1/\eta_{2-}(-q^{1/4}) + 1/\eta_{2-}(-iq^{1/4}) \\
\theta_{\Lambda_2+v_2}/\Delta(q) &= 1/\eta_{2-}(q^{1/4}) - 1/\eta_{2-}(iq^{1/4}) + 1/\eta_{2-}(-q^{1/4}) - 1/\eta_{2-}(-iq^{1/4})
\end{aligned}$$

Proof. If we multiply both sides by $\Delta(q)$ then both sides are modular forms of weight 8 and level 2, so it is sufficient to check that the coefficients of q^0, $q^{1/2}$, q^1, $q^{3/2}$, and q^2 of both sides are equal.

12 The fake monster Lie algebra.

In the next section we construct a family of superalgebras whose denominator formulas are twisted versions of the denominator formula of the fake monster Lie algebra, in the same way that we constructed a family of algebras from the monster Lie algebra. To do this we need a detailed description of the fake monster Lie algebra, which we give in this section.

As the Leech lattice Λ is an even lattice, it has a unique central extension

$$0 \to \{1, -1\} \to \hat{\Lambda} \to \Lambda \to 0$$

by a group of order 2, such that the commutator of any inverse images of $r, s \in \Lambda$ is $(-1)^{(r,s)}$. This central extension is unique up to nonunique isomorphism, and its group of automorphisms $\mathrm{Aut}(\hat{\Lambda})$ preserving the inner product is a nonsplit extension $2^{24}.\mathrm{Aut}(\Lambda)$. (The double cover of this nonsplit extension is isomorphic to the double cover of a centralizer of an involution of type 2B in the monster simple group.)

We start by recalling some results about the fake monster Lie algebra from [8] (where it is called the monster Lie algebra). The root lattice of the fake monster Lie algebra is the 26 dimensional even unimodular Lorentzian lattice $II_{25,1} = \Lambda \oplus II_{1,1}$, where the norm of an element $(\lambda, m, n) \in \Lambda \oplus II_{1,1}$ is defined to be $\lambda^2 - 2mn$. The real simple roots are the

norm 2 vectors of the form $(\lambda, 1, \lambda^2/2-1)$, and the imaginary simple roots are the positive multiples $n\rho$ of the Weyl vector $\rho = (0, 0, 1)$, each with multiplicity 24. A nonzero vector $r \in II_{25,1}$ is a root if and only if $r^2 \leq 2$, in which case it has multiplicity $p_{24}(1 - r^2/2)$ where $p_{24}(n)$ is the number of partitions of n into parts of 24 colours. The fake monster Lie algebra is essentially the space of physical vectors of the vertex algebra of $\hat{II}_{25,1}$, and can be constructed from the vertex algebra V_Λ of $\hat{\Lambda}$ in the same way that the monster Lie algebra M was constructed from the monster vertex algebra V in section 6.

The fake monster Lie algebra is acted on by the group of affine automorphisms of the Leech lattice which is a split extension $\Lambda.\mathrm{Aut}(\Lambda)$, because its Dynkin diagram may be interpreted as the Leech lattice as it has one real simple root $(\lambda, 1, \lambda^2/2 - 1)$ for each Leech lattice vector λ. However this is the wrong group of automorphisms to use. A more useful group of automorphisms of the fake monster Lie algebra is the nonsplit extension $2^{24}.\mathrm{Aut}(\Lambda) = \mathrm{Aut}(\hat{\Lambda})$, which acts naturally on the vertex algebra of the central extension $\hat{\Lambda}$ of the Leech lattice Λ, and therefore on the fake monster Lie algebra. The point is that we can use the no-ghost theorem 5.1 to describe how the group $\mathrm{Aut}(\hat{\Lambda})$ acts on the root spaces of the fake monster Lie algebra, and it is not obvious how we could do this if we used the group $\mathrm{Aut}(\Lambda)$.

We now describe what the fake monster Lie algebra M_Λ looks like as a module over $\mathrm{Aut}(\hat{\Lambda}) = 2^{24}\mathrm{Aut}(\Lambda)$. If we forget the Λ in the $II_{25,1} = \Lambda \oplus II_{1,1}$ grading of M_Λ then we can consider M_Λ to be a $II_{1,1}$-graded Lie algebra. The piece of M_Λ of degree $(m, n) \in II_{1,1}$ is then isomorphic as a Λ-graded $2^{24}\mathrm{Aut}(\Lambda)$ module to the weight $1 - mn$ piece of the vertex algebra V_Λ of $\hat{\Lambda}$ by theorem 6.1, so it is sufficient to describe V_Λ explicitly, which we can do as follows.

We let $S = S(\oplus_{i>0}\Lambda_i)$ be the symmetric algebra on the sum of a countable number of copies Λ_i of the Leech lattice Λ, and grade this by letting the elements of Λ_i have weight i. The dimension of the subspace of S of weight n is then equal to $p_{24}(n)$, the number of partitions of n into parts of $24 = \dim(\Lambda)$ colours, which is the coefficient of q^{n-1} of $1/\Delta(q) = q^{-1}\prod_{i>0}(1 - q^i)^{-24}$. We let $\mathbf{Z}(\hat{\Lambda})$ be the twisted group ring of Λ, which is the group ring of $\hat{\Lambda}$ quotiented out by the ideal $(1 + \epsilon)$ where ϵ is the element of the group ring of $\hat{\Lambda}$ corresponding to $-1 \in \hat{\Lambda}$, and grade it by letting e^v have weight $(v, v)/2$ for $v \in \hat{\Lambda}$. (This grading does not have the property that $\deg(ab) = \deg(a) + \deg(b)$.) Finally the graded space V_Λ is isomorphic to $\mathbf{Z}(\hat{\Lambda}) \otimes S$. The group $2^{24}\mathrm{Aut}(\Lambda) = \mathrm{Aut}(\hat{\Lambda})$ acts naturally on both $\hat{\Lambda}$ and on Λ_i, and so acts on V_Λ. Finally the Λ grading on V_Λ comes from the Λ grading on $\mathbf{Z}(\hat{\Lambda})$. (The Λ-grading on S is the trivial one such that everything has degree 0.) This gives a complete description of V_Λ as a Λ-graded $\mathrm{Aut}(\hat{\Lambda})$ module, and therefore describes M_Λ as a $II_{25,1}$-graded $\mathrm{Aut}(\hat{\Lambda})$ module.

We need a lemma to construct some elements of $\mathrm{Aut}(\hat{\Lambda})$ from elements of $\mathrm{Aut}(\Lambda)$.

Lemma 12.1. *Let σ be an automorphism of Λ of order n, let Λ^σ the vectors of Λ fixed by σ, and let $\hat{\Lambda}^\sigma$ be the inverse image of Λ^σ in $\hat{\Lambda}$. If n is odd then σ can be lifted to an element $\hat{\sigma} \in \mathrm{Aut}(\Lambda)$ of order n which fixes all elements of $\hat{\Lambda}^\sigma$. If n is even then σ can be lifted to an element $\hat{\sigma} \in \mathrm{Aut}(\Lambda)$ which fixes all elements of $\hat{\Lambda}^\sigma$. The automorphism $\hat{\sigma}^n$ multiplies any element v of $\hat{\Lambda}$ by $\epsilon^{(v,\sigma^{n/2}(v))}$, and in particular $\hat{\sigma}$ has order n if $(v, \sigma^{n/2}(v))$ is even for all v, and has order $2n$ otherwise.*

Proof. If n is odd this result is obvious because σ has lifts of the same odd order, and any lift of odd order must fix all elements of $\hat{\Lambda}^g$. If n is even then we can multiply any lift of σ by some element of 2^{24} to get an element $\hat{\sigma}$ which fixes all elements of $\hat{\Lambda}^\sigma$. The element $v\sigma(v)\dots\sigma^{n-1}(v)$ is fixed by σ as it is in $\hat{\Lambda}^\sigma$ and is therefore equal to $\sigma(v)\dots\sigma^n(v)$. Hence $v = \sigma^n(v)$ if and only if v commutes with $\sigma(v)\dots\sigma^{n-1}(v)$, which is true if and only if $(v,\sigma(v)) + \dots (v,\sigma^{n-1}(v))$ is even, which is true if and only if $(v,\sigma^{n/2}v)$ is even because $(v,\sigma^i(v)) = (v,\sigma^{n-i}(v))$. Therefore $\sigma^n(v)$ fixes v if and only if $(v,\sigma^{n/2}(v))$ is even, and therefore multiplies v by $\epsilon^{(v,\sigma^{n/2}(v))}$. This proves lemma 12.1.

For example, if σ is an involution of $\mathrm{Aut}(\Lambda)$ whose lattice Λ^σ has dimension 8 or 16 then it is easy to check that $(v,\sigma(v))$ is even for all $v \in \Lambda$, so σ can be lifted to an element $\hat{\sigma} \in \mathrm{Aut}(\hat{\Lambda})$ which has order 2 and fixes all elements of $\hat{\Lambda}^\sigma$. This is not possible if σ is an involution with Λ^σ 12-dimensional, because there are vectors v such that $(v,\sigma(v))$ is odd.

The homology $H(E)$ of the subalgebra E of M_Λ corresponding to the positive roots can be described as follows. We let e^ρ stand for a one dimensional $II_{25,1}$-graded vector space of degree $\rho \in II_{25,1}$. Then the virtual vector space $e^\rho H(E)$ is antisymmetric under the Weyl group, so it is only necessary to describe the subspace of it generated by elements whose degrees are in the Weyl chamber of $II_{25,1}$. By the remarks at the end of section 4 this subspace is naturally isomorphic to the virtual vector space $\Lambda(\oplus_{i>0}\Lambda_i)$, where Λ_i is a copy of the vector space of Λ with degree $i\rho \in II_{25,1}$. This follows from the fact that the imaginary simple roots of M_Λ are the positive multiples of ρ each with multiplicity 24, so all imaginary simple roots are orthogonal to each other, so the subspace of the homology group that we need is just the exterior power of the sum of the root spaces of the imaginary simple roots.

13 The denominator formula for fake monster Lie algebras.

We can now calculate a twisted denominator formula for each element of $\mathrm{Aut}(\hat{\Lambda})$, which is often the denominator formula for some other Lie algebra or superalgebra. In this section we calculate a few examples of these Lie algebras.

We let g be some element of $\mathrm{Aut}(\hat{\Lambda}) = 2^{24}\mathrm{Aut}(\Lambda)$ of order N, and we let L be the sublattice of Λ fixed by g. The projection of Λ into the vector space of L is the dual lattice L' of L because Λ is unimodular. For simplicity we assume that any power g^n of g fixes all elements of $\hat{\Lambda}$ which are in the inverse image of any vector of Λ fixed by g^n. (If we do not assume this condition we can still carry out most of the work here, but we would have to put more effort into keeping track of signs.) Lemma 12.1 provides us with plenty of examples of such elements g; for example any element of odd order.

We consider both sides of the formula

$$\Lambda(E) = H(E)$$

to be L'-graded virtual g modules. We calculate the trace of g on both sides, which can be thought of as an element of some completion of the group ring of L'. We can calculate the trace of g on $\Lambda(E)$ provided we know exactly how g acts on E, and we can calculate the trace of g on $H(E)$ if we know exactly how g acts on the simple root spaces of M_Λ.

We first calculate the trace of g on $\Lambda(E) = \exp(-\sum_{i>0} \psi^i(E)/i)$ (where ψ^i are the Adams operations). This trace is equal to

$$\exp(-\sum_{v\in L'}\sum_{i>0} \mathrm{Tr}(g^i|E_v)e^{iv}/i) \tag{13.1}$$

where E_v is the subspace of E whose L'-degree is v. (Note added May 1999: the published version of this formula incorrectly had a term E_{iv} instead of E_v.) We wish to express the trace of g in the form $\prod(1-e^r)^{\mathrm{mult}(r)} = \exp -\sum_r \sum_{i>0} e^{ir}\mathrm{mult}(r)/i$ for some numbers $\mathrm{mult}(r)$. By applying the Moebius inversion formula to (13.1) as in section 11 we see that the numbers $\mathrm{mult}(r)$ are given by

$$\mathrm{mult}(r) = \sum_{ds|((r,L),N)} \mu(s)\mathrm{Tr}(g^d|E_{r/ds})/ds \tag{13.2}$$

where (r, L) is the highest common factor of the numbers (r, a) for $a \in L$.

By theorem 2.2 of [6] there is a reflection group W^g acting on the lattice L, which may be taken as any of the following groups.

1 The subgroup of W of elements that commute with g.
2 The subgroup of W of elements that map ρ into L.
3 The subgroup of W of elements mapping L into L.

The positive roots of W^g are the vectors which are the sums of the conjugates of some positive real root of $II_{25,1}$. The simple roots of W^g are the sums of orbits of simple roots of W that have positive norm, and they are also the roots of W^g satisfying $(r, \rho) = -r^2/2$, so ρ is a norm 0 Weyl vector for W^g. (Warning: W^g is not always the full reflection group of L.)

From (13.2) we can see that any 2 positive vectors of L' conjugate under W^g have the same multiplicity. If r is a real simple root of W^g then an easy calculation using (13.2) shows that $\mathrm{mult}(r)$ is 1, so $\mathrm{mult}(r)$ is 1 for any positive real root of W^g, because any such root is conjugate under W^g to a simple root of W^g. If r has positive norm but is not a root of W^g then $\mathrm{mult}(r) = 0$.

The product

$$e^\rho \prod_{r\in L}(1-e^r)^{\mathrm{mult}(r)}$$

is therefore antisymmetric under the reflection group W^g, so it is equal to $\sum_{w\in W^g} \det(w)w(e^\rho S)$ for some S with $e^\rho S$ in the Weyl chamber of W^g. (The determinant of w means its determinant as an automorphism of L, rather than as an automorphism of Λ.) This sum is also equal to the trace of g on $e^\rho H(E)$ because it is equal to the trace of g on $e^\rho\Lambda(E)$, so S is the trace of g on the subspace of $H(E)$ in the Weyl chamber of W^g. This subspace is isomorphic to $\Lambda(\oplus_{i>0}\Lambda_i)$ where Λ_i is a copy of the Leech lattice with degree $i\rho \in II_{25,1}$. The trace of g on this subspace is just $e^{-\rho}\eta_g(e^\rho)$ where $\eta_g(q)$ is defined to be $\eta(\epsilon_1 q)\ldots\eta(\epsilon_{24}q)$ if g has eigenvalues ϵ_1, ... , ϵ_{24} on the vector space of Λ. (If G has generalized cycle shape $a_1^{b_1}a_2^{b_2}\ldots$ then $\eta_g(q) = \eta(q^{a_1})^{b_1}\eta(q^{a_2})^{b_2}\ldots$.) Therefore we finally get

$$e^\rho \prod_{r\in L}(1-e^r)^{\mathrm{mult}(r)} = \sum_{w\in W^g} \det(w)w(\eta_g(e^\rho)). \tag{13.3}$$

The right hand side of this is exactly the denominator formula for a generalized Kac-Moody superalgebra with the following simple roots.

1 The real simple roots are the simple roots of the reflection group W^g, which are the roots r with $(r,\rho) = -(r,r)/2$.

2 The imaginary simple roots are the positive multiples $i\rho$ of the Weyl vector ρ, with multiplicity equal to

$$\text{mult}(i\rho) = \sum_{ja_k=i} b_k$$

if g has generalized cycle shape $a_1^{b_1}a_2^{b_2}\dots$.

Therefore the left hand side is the left hand side of the denominator formula for this generalized Kac-Moody superalgebra, so its positive roots have multiplicity mult(r).

14 Examples of fake monster Lie algebras.

We now evaluate both sides of the denominator formula (13.3) for several elements of $\text{Aut}(\hat\Lambda)$, so we get several example of generalized Kac-Moody algebras whose simple roots and root multiplicities are known explicitly.

Example 1. The fake baby monster Lie algebra of rank 18. For this example we let $g \in \text{Aut}(\hat\Lambda)$ be an element of order 2 which is the lift of an element of order 2 of $\text{Aut}(\Lambda)$ which fixes a 16 dimensional sublattice Λ^g of Λ, such that g fixes all elements of $\hat\Lambda^g$. By lemma 12.1 such an element g exists, and the lattice Λ^g is the Barnes-Wall lattice of dimension 16 and determinant 2^8 [12]. The root lattice L is the sum of Λ^g and the two dimensional even Lorentzian lattice $II_{1,1}$. The Barnes-Wall lattice Λ^g has no roots, so by theorem 3.1 of [6] the group W^g is the full reflection group of L.

We can calculate the multiplicities mult(r) explicitly. The result we get is that mult$(r) = p_g(1-r^2/2)$ if $r \in L$ and $r \notin 2L'$, mult$(r) = p_g(1-r^2/2) + p_g(1-r^2/4)$ if $r \in 2L'$, and mult$(r) = 0$ otherwise, where $p_g(1+n)$ is the coefficient of q^n in $1/\eta_g(q) = q^{-1}\prod_{i>0}(1-q^i)^{-8}(1-q^{2i})^{-8} = q^{-1} + 8 + 44q + \dots$. If $r \in L'$ then $\text{Tr}(g|E_r)$ is 0 if $r \notin L$ and equal to the coefficient of $q^{-r^2/2}$ in $1/\eta_g(q)$ if $r \in L$. The value of $\text{Tr}(g^2|E_r)$ is just the dimension of E_r, which is the coefficient of $q^{-r^2/2}$ in $\theta_{\Lambda^{g\perp}+r^\perp}(q)/\Delta(q)$ where $\theta_{\Lambda^{g\perp}+r^\perp}(q)$ is the theta function of the coset $\Lambda^{g\perp}+r^\perp$ of the lattice $\Lambda^{g\perp}$ and where $r^\perp$ is a vector such that $r+r^\perp$ is in $II_{25,1}$. If we substitute these values into the formula

$$\text{mult}(r) = \text{Tr}(g|E_r) - \text{Tr}(g|E_{r/2})/2 + \text{Tr}(g^2|E_{r/2})/2$$

we find that mult(r) is equal to the coefficient of $q^{-r^2/2}$ in

$$\begin{aligned} 0 &\text{ if } r \notin L \\ 1/\eta_g(q) &\text{ if } r \in L, r \notin 2L' \\ 1/\eta_g(q) + \theta_{\Lambda^{g\perp}+r^\perp}(q^4)/\Delta(q^4) &\text{ if } r \in 2L', r \notin 2L \\ 1/\eta_g(q) - 1/\eta_g(q^4)/2 + \theta_{\Lambda^{g\perp}}(q^4)/\Delta(q^4)/2 &\text{ if } r \in 2L. \end{aligned}$$

By lemma 11.1 this implies that if $r \in 2L'$ then mult(r) is equal to the coefficient of $q^{-r^2/2}$ of $1/\eta_g(q) + 1/\eta_g(q^2)$. Therefore the explicit version of the product formula for the fake

baby monster Lie algebra is

$$e^\rho \prod_{r\in L^+}(1-e^r)^{p_g(1-r^2/2)} \prod_{r\in 2L'^+}(1-e^r)^{p_g(1-r^2/4)} = \sum_{w\in W^g} \det(w) w(e^\rho \prod_{i>0}(1-e^{i\rho})^8(1-e^{2i\rho})^8)$$

where L is the Lorentzian lattice which is the sum of the Barnes-Wall lattice Λ^g and the two dimensional even Lorentzian lattice, W^g is its reflection group which has Weyl vector ρ, and L' is the dual of L.

Example 2. We can construct a similar Lie algebra for any of the primes $p = 2$, 3, 5, 7, 11, or 23 with $(p+1)|24$. We let g be an element of $\mathrm{Aut}(\hat\Lambda)$ of order p corresponding to an element of $M_{24} \subset \mathrm{Aut}(\Lambda)$ of cycle shape $1^{24/(p+1)}p^{24/(p+1)}$. We assume that the lattice Λ^g of dimension $48/(p+1)$ has no roots and if $r \in \Lambda^{g\perp\prime}$ then

$$\theta_{\Lambda^{g\perp}+r}(q) = \eta(q)^{24p/(p+1)}(\eta(q^p)^{-24/(p+1)}\delta(r\in\Lambda^{g\perp}) + \sum_{0\le i<p} \epsilon^{-ipr^2/2}\eta(\epsilon^i q^{1/p})^{-24/(p+1)})$$

where ϵ is a primitive p'th root of 1 and where $\delta(r \in \Lambda^{g\perp})$ is 1 if $r \in \Lambda^{g\perp}$ and 0 otherwise. These assumptions are in principle not difficult to prove if they are true, as they are essentially just identities between modular forms. (For $p = 2$ they are just lemma 12.1.)

Assuming these identities, an argument similar to the one above for $p = 2$ shows that the denominator formula for the Lie algebra of G is

$$\begin{aligned} &e^\rho \prod_{r\in L^+}(1-e^r)^{p_g(1-r^2/2)} \prod_{r\in pL'^+}(1-e^r)^{p_g(1-r^2/2p)} \\ &= \sum_{w\in W^g} \det(w) w(e^\rho \prod_{i>0}(1-e^{i\rho})^{24/(p+1)}(1-e^{pi\rho})^{24/(p+1)}) \end{aligned}$$

where $\sum_{i>0} p_g(1+i)q^i = 1/\eta_g(q)$. For p= 2, 3, 5, 7, and 11 these Lie algebras seem to correspond to the baby monster, the Fischer group Fi_{24}, the Harada Norton group, the Held group and the Mathieu group M_{12} in the same way that the fake monster Lie algebra corresponds to the monster simple group.

Example 3. The fake Conway Lie superalgebra of rank 10. This is the algebra described at the end of section 2, and seems to correspond to Conway's simple group Co_1. We let $g \in \mathrm{Aut}(\hat\Lambda)$ be an element of order 2 which is the lift of an element of order 2 of $\mathrm{Aut}(\Lambda)$ which fixes a lattice Λ^g of Λ of dimension 8. This lattice Λ^g is isomorphic to the E_8 lattice with all norms doubled, so if we halve all the norms of the lattice $\Lambda^g \oplus II_{1,1}$ we get the nonintegral lattice of determinant 1/4 which is the dual of the sublattice of even vectors of $I_{9,1}$. A calculation similar to that in example 1 but using lemma 11.2 shows that the denominator formula of this Lie superalgebra is

$$e^\rho \prod_{r\in\Pi^+}(1-e^r)^{\mathrm{mult}(r)} = \sum_{w\in W} \det(w) w(e^\rho \prod_{n>0}(1-e^{n\rho})^{(-1)^n 8})$$

where the multiplicity of the root $r = (v, m, n) \in L$ is equal to

$$\mathrm{mult}(r) = (-1)^{(m-1)(n-1)}p_g((1-r^2)/2) = (-1)^{m+n}|p_g((1-r^2)/2)|,$$

and $p_g(n)$ is defined by

$$\sum p_g(n)q^n = q^{-1/2}\prod_{n>0}(1-q^{n/2})^{-(-1)^n 8}.$$

15 Open questions

We list a few conjectures and open questions about the Lie algebras and superalgebras we have constructed.

(1) Prove the assumptions about the modular forms used in example 2 in section 14. In principle this should be easy (at least when they are true) because they just involve checking a finite number of identities between modular forms.

(2) Investigate the Lie algebras and superalgebras coming from other elements of the monster or $\mathrm{Aut}(\Lambda)$ and write down their denominator formulas explicitly in some nice form.

(3) Find a natural construction for these Lie algebras and superalgebras (i.e. other than by generators and relations). We used natural constructions for the monster Lie algebra and the fake monster Lie algebra from vertex algebras, but I do not know of any similar constructions for most of the other Lie algebras. The easiest case is the superalgebra of rank 10 which can be constructed from superstrings on a 10-dimensional torus. (The even part of the rank 10 superalgebra was constructed in [3, 4]; the odd part is more difficult to construct.) A natural construction should give actions of various finite groups on these Lie algebras; for example the double cover of the baby monster should act on the baby monster Lie algebra.

(4) Describe the Lie bracket from $V_{ab}\otimes V_{cd}$ to $V_{(a+c)(b+d)}$ of the monster Lie algebra explicitly in terms of the vertex algebra operations on V.

(5) Is the baby monster Lie algebra a subalgebra of the monster Lie algebra in a way that preserves the action of the double cover of the baby monster? Similarly for the other monstrous Lie algebras.

(6) Are there any generalized Kac-Moody algebras, other than the finite dimensional, affine, monstrous or fake monstrous ones, whose simple roots and root multiplicities can both be described explicitly? The monstrous and fake monstrous algebras are both finite families, each with a few hundred members, corresponding roughly to the conjugacy classes in the monster and in $\mathrm{Aut}(\hat{\Lambda})$. One example of a denominator formula for a Lie superalgebra of rank 3 is the identity

$$\sum_{i+j+k=0}(-p)^{jk}(-q)^{ik}(-r)^{ij} = \prod_{i+j+k>0}((1-p^iq^jr^k)/(1+p^iq^jr^k))^{c(ij+jk+ki)}$$

where $c(i)$ is defined by $\sum_n c(n)q^n = \prod_{n>0}(1+q^n)/(1-q^n) = 1+2q+4q^2+8q^3+14q^4+\ldots$.

(7) Find all completely replicable functions. (The ones with integral coefficients were found by computer in [1].) Are nontrivial completely replicable functions always modular functions of genus 0? A proof of this conjecture of Norton's [28] which was not a case by case verification would be much neater than the argument in section 9.

(8) Is it possible to say anything interesting from Lie algebras constructed from the vertex algebras of lattices (other than the Leech lattice) as in section 6? The two obvious

candidates are the lattices E_8 and $E_8 \oplus E_8$, so that the corresponding Lie algebras have root lattices $II_{9,1}$ and $II_{17,1}$. The real simple roots of these Lie algebras are the Dynkin diagrams of the reflection groups of the lattices, which were described by Vinberg. There are some calculations connected with the Lie algebra of $II_{9,1}$ in [24].

References

[1]. D. Alexander, C. Cummins, J. McKay, C. Simons, Completely replicable functions, preprint.

[2]. M. F. Atiyah, K-theory, Benjamin, Inc. 1967.

[3]. R. E. Borcherds, Vertex algebras, Kac-Moody algebras, and the monster. Proc. Natl. Acad. Sci. USA. Vol. 83 (1986) 3068-3071.

[4]. R. E. Borcherds, Generalized Kac-Moody algebras. J. Algebra 115 (1988), 501–512.

[5]. R. E. Borcherds, Central extensions of generalized Kac-Moody algebras. J. Alg. 140, 330-335 (1991).

[6]. R. E. Borcherds, Lattices like the Leech lattice, J. Algebra, vol 130, No. 1, April 1990, 219-234.

[7]. R. E. Borcherds, J. H. Conway, L. Queen, N. J. A. Sloane, A monster Lie algebra?, Adv. Math. 53 (1984) 75-79. This paper is reprinted as chapter 30 of [12].

[8]. R. E. Borcherds, The monster Lie algebra, Adv. Math. Vol. 83, No. 1, Sept. 1990.

[9]. R. E. Borcherds, Vertex algebras, to appear.

[10]. H. Cartan, S. Eilenberg, Homological algebra, Princeton University Press 1956.

[11]. J. H. Conway, The automorphism group of the 26 dimensional even Lorentzian lattice. J. Algebra 80 (1983) 159-163. This paper is reprinted as chapter 27 of [12].

[12]. J. H. Conway, N. J. A. Sloane. Sphere packings, lattices and groups. Springer-Verlag New York 1988, Grundlehren der mathematischen Wissenschaften 290.

[13]. J. H. Conway, S. Norton, Monstrous moonshine, Bull. London. Math. Soc. 11 (1979) 308-339.

[14]. J. H. Conway, R. T. Curtis, S. P. Norton, R. A. Parker, R. A. Wilson, Atlas of finite groups, Clarendon Press, Oxford, 1985.

[15]. I. B. Frenkel, Representations of Kac-Moody algebras and dual resonance models, Applications of group theory in theoretical physics, Lect. Appl. Math. 21, A.M.S. (1985), p.325-353.

[16]. I. B. Frenkel, J. Lepowsky, A. Meurman, Vertex operator algebras and the monster, Academic press 1988.

[17]. I. B. Frenkel, J. Lepowsky, A. Meurman, A natural representation of the Fischer-Griess monster with the modular function J as character, Proc. Natl. Acad. Sci. USA 81 (1984), 3256-3260.

[18]. I. B. Frenkel, Y-Z. Huang, J. Lepowsky, On axiomatic formulations of vertex operator algebras and modules, preprint.

[19]. I. B. Frenkel, H. Garland, G. Zuckerman, Semi-infinite cohomology and string theory, Proc. Natl. Acad. Sci. U.S.A. 83 (1986), 8442-8446.

[20]. H. Garland and J. Lepowsky, Lie algebra homology and the Macdonald-Kac formulas, Invent. math. 34 (1976), p.37-76.

[21]. P. Goddard and C. B. Thorn, Compatibility of the dual Pomeron with unitarity and the absence of ghosts in the dual resonance model, Phys. Lett., B 40, No. 2 (1972), 235-238.
[22]. R. C. Gunning, Lectures on modular forms, Annals of mathematical studies, Princeton University Press, 1962.
[23]. V. G. Kac, "Infinite dimensional Lie algebras", third edition, Cambridge University Press, 1990. (The first and second editions (Birkhauser, Basel, 1983, and C.U.P., 1985) do not contain the material on generalized Kac-Moody algebras that we need.)
[24]. V. G. Kac, R. V. Moody, M. Wakimoto, On E_{10}, preprint.
[25]. M. Koike, On Replication Formula and Hecke Operators, Nagoya University preprint.
[26]. B. Kostant, Lie algebra cohomology and the generalized Borel-Weil theorem. Annals of Math. 74, 329-387 (1961).
[27]. K. Mahler, On a class of non-linear functional equations connected with modular functions, J. Austral. Math. Soc. 22A, (1976), 65-118.
[28]. S. P. Norton, More on moonshine, Computational group theory, Academic press, 1984, 185-193.
[29]. S. P. Norton, Generalized Moonshine, Proc. Symp. Pure Math. 47 (1987) p. 208-209.
[30]. J. P. Serre, A course in arithmetic. Graduate texts in mathematics 7, Springer-Verlag, 1973.
[31]. J. G. Thompson, A finiteness theorem for subgroups of $PSL(2, \mathbf{R})$ which are commensurable with $PSL(2, \mathbf{Z})$. Proc. Symp. Pure Math. 37 (1979) p. 533-555.

Séminaire Delange-Pisot-Poitou. Théorie des nombres

ANDREW P. OGG

Automorphismes de courbes modulaires

Séminaire Delange-Pisot-Poitou. Théorie des nombres, tome 16, n° 1 (1974-1975), exp. n° 7, p. 1-8

<http://www.numdam.org/item?id=SDPP_1974-1975__16_1_A4_0>

© Séminaire Delange-Pisot-Poitou. Théorie des nombres
(Secrétariat mathématique, Paris), 1974-1975, tous droits réservés.

L'accès aux archives de la collection « Séminaire Delange-Pisot-Poitou. Théorie des nombres » implique l'accord avec les conditions générales d'utilisation (http://www.numdam.org/conditions). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM
Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques
http://www.numdam.org/

Séminaire DELANGE-PISOT-POITOU
(Théorie des nombres)
16e année, 1974/75, n° 7, 8 p.

9 décembre 1974

AUTOMORPHISMES DE COURBES MODULAIRES

par Andrew P. OGG

Soient N un entier positif, et $\Gamma_0(N)$ le sous-groupe du groupe modulaire $\Gamma = SL(2, \underline{Z})/\pm 1$ défini par les matrices $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ avec N divisant c. Alors $\Gamma_0(N)$ agit sur le demi-plan supérieur $\mathfrak{H} = \{\tau = x + iy ; y > 0\}$ par $\tau \longmapsto \frac{a\tau+b}{c\tau+d}$.

Soient $Y_0(N) = \Gamma_0(N) \backslash \mathfrak{H}$, et $X_0(N)$ la compactification de $Y_0(N)$ par adjonction des pointes (points paraboliques). Alors $X_0(N)$ est une courbe (projective, non singulière) définie sur $\underline{Q}$; son genre $g = g(N)$ est donné par une formule connue. Par exemple, si N est premier, alors $X_0(N)$ a exactement deux pointes, 0 et ∞, et elles sont rationnelles sur $\underline{Q}$; le genre est 0 si $N = 2$ ou 3, et

$$(1) \qquad g = \frac{N - a}{12} \quad \text{si} \quad \begin{cases} N \equiv a \pmod{12} \\ a = -1, 5, 7, 13. \end{cases}$$

Si u est une involution (automorphisme d'ordre 2) de $X_0(N)$, le genre du quotient $X_0(N)/(u)$ est

$$(2) \qquad g^{(u)} = \frac{g + 1}{2} - \frac{n(u)}{4},$$

d'après la formule de Riemann-Hurwitz, où $n(u)$ est le nombre des points fixes de u. En particulier, $X_0(N)$ a l'involution $w = w_N$ définie par la matrice $\begin{pmatrix} 0 & -1 \\ N & 0 \end{pmatrix}$, pour laquelle $n(w)$ est donné par une formule de Fricke :

$$(3) \qquad n(w) = \begin{cases} h(-N) + h(-4N) & (\text{si } N \equiv 3 \pmod 4) \\ h(-4N) & (\text{sinon}), \end{cases}$$

où $h(-m)$ est le nombre des classes des formes quadratiques binaires positives primitives de discriminant $-m$. On pose $X_0^+(N) = X_0(N)/(w)$ et

$$g^+ = g^{(w)} = \text{genre de } X_0^+(N).$$

L'involution w est définie sur $\underline{Q}$, et $X_0^+(N)$ est définie sur $\underline{Q}$.

Si $g \geqslant 2$, alors le groupe $A = \mathrm{Aut}(X_0(N))$ des automorphismes de $X_0(N)$ est fini. Il contient le sous-groupe B des automorphismes "modulaires", i. e. $B = \mathfrak{N}/\Gamma_0(N)$, où $\mathfrak{N}$ est le normalisateur de $\Gamma_0(N)$ dans $\mathrm{Aut}(\mathfrak{H}) = SL(2, \underline{R})/\pm 1$. LEHNER et NEWMAN [2] ont déterminé le groupe B. "En général", i. e. si $4 \nmid N$ et $9 \nmid N$, le groupe B est le groupe W des involutions partielles d'Atkin-Lehner, isomorphe au produit de n groupes d'ordre 2, où n est le nombre des nombres premiers distincts qui divisent N. Les éléments de W sont définis sur $\underline{Q}$. En particulier, $B = \{1, w\}$ si N est premier. Un élément de $A - B$ sera appelé "exceptionnel" ; le seul exemple connu où $A - B$ n'est pas vide est le cas $N=37$.

Avant de discuter cet exemple, rappelons qu'une courbe X, de genre ≥ 2, est hyperelliptique si elle possède une fonction $f : X \longrightarrow \mathbb{P}^1$ de degré 2, i. e. si elle possède une involution hyperelliptique v telle que $X/(v)$ est de genre 0. Dans ce cas, v est unique (par exemple parce qu'elle induit -1 sur la jacobienne de X, ou sur l'espace des différentielles holomorphes sur X) ; donc v est dans le centre de $\mathrm{Aut}(X)$, et définie sur tout corps de définition de X. Dans la suite, la lettre v est toujours réservée aux involutions hyperelliptiques, et la lettre w à $\begin{pmatrix} 0 & -1 \\ N & 0 \end{pmatrix}$. Enfin, toute courbe de genre 2 est hyperelliptique.

Pour $N = 37$, on a $g = 2$ et $g^+ = 1$, et donc $X_0(37)$ est hyperelliptique avec $v \neq w$ (i. e. v est exceptionnelle). L'involution v applique les deux pointes de $X_0(37)$ sur deux points rationnels de $Y_0(37)$, en fait les seuls éléments rationnels de $Y_0(37)$ (cf. [5]). Le problème des automorphismes exceptionnels se trouve ainsi lié au problème des points rationnels de $Y_0(N)$ (cf. [7]).

THÉORÈME 1 [6]. - $X_0(N)$ est hyperelliptique pour exactement dix-neuf valeurs de N. Le seul cas où v est exceptionnelle est $N = 37$. Les autres dix-huit valeurs sont $N = 22, 23, 26, 28, 29, 30, 31, 33, 35, 39, 40, 41, 46, 47, 48, 50, 59, 71$.

Le vrai problème (évidemment plus difficile que celui résolu par le théorème 1) est de déterminer le groupe $A = \mathrm{Aut}(X_0(N))$ pour tout N (avec $g \geq 2$) ; pour l'instant, on ne connaît la réponse que dans le cas où N est premier.

THÉORÈME 2. - Si N est premier, $N \geq 23$ (i. e. $g \geq 2$), et $N \neq 37$, alors $A = \{1, w\}$. Pour $N = 37$, $A = \{1, w, v, vw\}$ est le groupe de Klein.

Le paragraphe 1 de cet article esquisse la démonstration du théorème 1 (sans détails, [6] étant maintenant disponible). Le paragraphe 2 donne la démonstration du théorème 2, complète si on admet des résultats inédits de K. RIBET et B. MAZUR ; des conversations avec A. BRUMER et J. VÉLU, en automne 1974, m'ont beaucoup aidé. Enfin, le paragraphe 3 discute le lien avec les courbes elliptiques supersingulières.

1. Le problème hyperelliptique (cf. [6]).

On aura besoin de l'interprétation modulaire (i. e. en termes de courbes elliptiques variables) de $X_0(N)$, qui est la suivante : On associe à $\tau \in \mathfrak{H}$ la paire (E, C), où $E = \mathbb{C}/L$ est la courbe elliptique correspondant au réseau $L = \mathbb{Z}\tau \oplus \mathbb{Z}$, et C est le sous-groupe cyclique de E, d'ordre N, engendré par l'image de $1/N$. Les paires (E, C) et (E', C'), associées à $\tau, \tau' \in \mathfrak{H}$, sont isomorphes (i. e. il existe un isomorphisme $f : E \xrightarrow{\sim} E'$ avec $f(C) = C'$) si, et seulement si, τ et τ' sont $\Gamma_0(N)$-équivalents. Il y a donc une bijection entre $Y_0(N)$ et les classes d'isomorphisme des paires (E, C), où E est une courbe elliptique, et C un sous-groupe cyclique d'ordre N. L'involution w associe à

la (classe de la) paire (E , C) la paire (E' , C') , où $E' = E/C$ et $C' = E_N/C$, $E_N = \{P \in E : N.P = 0\}$.

Si on préfère parler d'isogénies au lieu de sous-groupes cycliques, on a une isogénie $\lambda : E \longrightarrow E'$ de courbes elliptiques, dont le noyau C est cyclique d'ordre N , et $w(\lambda) = \hat{\lambda}$ est l'isogénie $E' \longrightarrow E$ avec $\hat{\lambda} \circ \lambda = N : E \longrightarrow E$ et $\lambda \circ \hat{\lambda} = N : E' \longrightarrow E'$.

Si ℓ est un nombre premier qui ne divise pas N , alors la courbe $X_0(N)$ a une bonne réduction modulo ℓ , encore notée $X_0(N)$, d'après un théorème d'Igusa. La courbe réduite $X_0(N)$ est donc une courbe non singulière sur $\underset{\sim}{F}_\ell$, du même genre g ; elle a le "même" ensemble de "pointes" $X_0(N) - Y_0(N)$ qu'en caractéristique 0 , et les éléments de $Y_0(N)$ correspondent aux classes de paires (E , C) , de courbes elliptiques E et sous-groupes cycliques C , en caractéristique ℓ .

Supposons maintenant $X_0(N)$ hyperelliptique, avec N impair ; alors la courbe réduite modulo 2 est encore hyperelliptique en caractéristique 2 . La courbe étant un revêtement double de $\underset{\sim}{P}^1$, le nombre de points rationnels dans un corps fini $\underset{\sim}{F}_{2^n}$ est $\leqslant 2(1 + 2^n)$; en particulier

$$|X_0(N)(\underset{\sim}{F}_4)| \leqslant 10 . \tag{4}$$

Considérons la courbe elliptique $E : y^2 + y = X^3$. Elle est la seule courbe elliptique supersingulière (i. e. $E_2 = 0$) en caractéristique 2 , et le groupe de ses automorphismes est d'ordre 24 . Sur $\underset{\sim}{F}_4$, le Frobenius $\varphi(x,y) = (x^4,y^4)$ est $\varphi = -2$, un entier, donc tous les sous-groupes finis C de E sont $\underset{\sim}{F}_4$-rationnels (i. e. stables sous l'action de $\mathrm{Gal}(\overline{\underset{\sim}{F}}_4/\underset{\sim}{F}_4)$, i. e. stables par φ). Le nombre des sous-groupes cycliques C de E , d'ordre N , est

$$\psi(N) = (\Gamma : \Gamma_0(N)) = N \prod_{p|N}(1 + 1/p) . \tag{5}$$

On a donc au moins $\psi(N)/12$ points de $X_0(N)(\underset{\sim}{F}_4)$ correspondants, parce qu'une paire (E , C) rationnelle sur un corps K représente un élément de $Y_0(N)(K)$ (cf. [1], [7]), et parce que $|\mathrm{Aut}(E)| = 24$, et l'automorphisme -1 de E fixe toute paire (E , C) . On a aussi les deux "pointes" 0 et ∞ , rationnelles sur $\underset{\sim}{F}_2$; en utilisant (4), on trouve $2 + \psi(N)/12 \leqslant 10$, i. e. $\psi(N) \leqslant 96$. Il y a des bornes analogues si $X_0(N)$ est hyperelliptique, et $p \nmid N$, pour $p = 3,5,7,\ldots$, donc avec l'aide de diverses astuces [6], on a le théorème 1.

Ce qui est agréable avec cette méthode, c'est qu'elle donne des majorations des valeurs de N pour lesquelles $X_0(N)$ est hyperelliptique, sans se soucier de savoir si v est exceptionnelle ou non.

2. Le groupe des automorphismes (N premier).

Soit $J = J_0(N)$ la jacobienne de $X = X_0(N)$; J est une variété abélienne sur $\underset{\sim}{Q}$ de dimension g . Soient $\mathrm{End}(J)$ l'anneau des endomorphismes de J , et $\mathrm{End}_{\underset{\sim}{Q}}(J)$ le sous-anneau des endomorphismes définis sur $\underset{\sim}{Q}$. Alors $\mathrm{End}_{\underset{\sim}{Q}}(J)$ contient l'algè-

bre de Hecke $\mathcal{H}$, engendrée par les opérateurs de Hecke.

L'algèbre $\mathcal{H} \otimes \underset{\sim}{Q}$ est commutative, et de dimension g sur $\underset{\sim}{Q}$; on a

$$\mathcal{H} \otimes \underset{\sim}{Q} \simeq K_1 \times \ldots \times K_n , \tag{6}$$

où K_i est un corps totalement réel. La décomposition (6) de $\mathcal{H} \otimes \underset{\sim}{Q}$ correspond à une décomposition de J, à isogénie près :

$$J \sim J^{(1)} \times \ldots \times J^{(n)} , \tag{7}$$

où $J^{(i)}$ est une variété abélienne, définie sur $\underset{\sim}{Q}$, et $\underset{\sim}{Q}$-simple, avec

$$\dim J^{(i)} = [K_i : \underset{\sim}{Q}] \quad \text{et} \quad \mathrm{End}_{\underset{\sim}{Q}}(J^{(i)}) \otimes \underset{\sim}{Q} = K_i .$$

K. RIBET a démontré que tout endomorphisme de J est défini sur $\underset{\sim}{Q}$, et est donc dans $\mathcal{H} \otimes \underset{\sim}{Q}$, par la "congruence d'Eichler-Shimura" :

$$\mathrm{End}(J) \otimes \underset{\sim}{Q} = \mathcal{H} \otimes \underset{\sim}{Q} . \tag{8}$$

[Plus généralement, RIBET montre qu'un endomorphisme d'une variété abélienne, "semi-stable" sur K, est défini sur une extension non ramifiée K' de K ; pour $K = \underset{\sim}{Q}$, on a nécessairement $K' = \underset{\sim}{Q}$. La démonstration, très simple, n'occupe qu'une page [8]. Ici "semi-stable" veut dire que la réduction, modulo une place quelconque, ne contient pas de groupes additifs ; dans notre cas, J a bonne réduction modulo ℓ pour $\ell \nmid N$ (IGUSA) et réduction multiplicative modulo N (DELIGNE).] En particulier, les facteurs $J^{(i)}$ de J sont absolument simples, et absolument non-isogènes.

Supposons maintenant $N \geq 23$, i. e. $g \geq 2$ (N étant toujours premier). Le groupe $A = \mathrm{Aut}(X)$ est fini, et continu dans $\mathrm{Aut}(J)$; les corps K_i étant réels, il s'ensuit de (6) et (8) que

$$A \subset (\pm 1) \times (\pm 1) \times \ldots \times (\pm 1) \quad (\text{n fois}). \tag{9}$$

En particulier, A est commutatif, et tout $u \in A$ est défini sur $\underset{\sim}{Q}$ et d'ordre ≤ 2.

Pour $N = 37$, $g = 2$, et J est isogène au produit de deux courbes elliptiques, d'où $A \subset (\pm 1) \times (\pm 1)$. Nous connaissons déjà quatre éléments de A, d'où la partie du théorème 2 qui concerne 37. Supposons $N \neq 37$ dorénavant.

LEMME. - Si $u \in A$ fixe la pointe ∞, alors $u = 1$.

Démonstration. - Les valeurs propres de u sur l'espace Ω des différentielles holomorphes sur X sont toutes ± 1. Si elles sont toutes -1, alors u est une involution hyperelliptique, donc $u = w$ d'après le théorème 1 ($N \neq 37$) en contradiction avec l'hypothèse $u(\infty) = \infty$ (parce que $\omega(\infty) = 0$). Supposons que u a les valeurs propres $+1$ et -1. Alors on peut choisir deux différentielles ω_1, $\omega_2 \in \Omega$ avec

$$\omega_1 \circ u = \omega_1 , \quad \omega_2 \circ u = -\omega_2 , \tag{10}$$

qui soient fonctions propres de $\mathcal{H}$. Le développement à la pointe ∞ de ω_i est donc de la forme

$$\omega_i = (\sum_{m=1}^{\infty} a_m q^m).\frac{dq}{q} \qquad (q = \exp(2\pi i\tau))$$

où $\sum_{m=i}^{\infty} a_m m^{-s}$ a un produit eulérien et en particulier $a_1 \neq 0$, disons $a_1 = 1$. Alors $\omega = \omega_1 + \omega_2$ ne s'annule pas en ∞ , mais $\omega \circ u = \omega_1 - \omega_2$ s'annule en ∞ , en contradiction avec $u(\infty) = \infty$. Les valeurs propres de u sont donc toutes $+ 1$, i. e. $u = 1$.

Q. E. D.

Le sous-groupe Z_n de $J(\underset{\sim}{Q})$, engendré par la classe du diviseur $(0) - (\infty)$, de degré 0 , est cyclique d'ordre

(11) $$n = \text{ numérateur de } \frac{N-1}{12} .$$

Tout récemment MAZUR ([3], [4]) a démontré qu'en fait Z_n est le groupe de tous les points rationnels sur J , d'ordre fini :

(12) $$Z_n = J(\underset{\sim}{Q})^{tors} \qquad (\text{"conjecture d'Ogg"}).$$

Démontrons maintenant le théorème 2. Si $u \in A$, alors u agit sur le groupe Z_n (cf. (12)) ; on a donc

(13) $$(u0) - (u\infty) \sim m((0) - (\infty)) ,$$

où $m \in \underset{\sim}{Z}/n\underset{\sim}{Z}$.

Si $m = 1$, alors $(u0) + (\infty) \sim (0) + (u\infty)$. Si $u\infty = \infty$ alors $u = 1$, d'après le lemme. Si $u\infty \neq \infty$, alors X est hyperelliptique (parce que $u\infty \neq u0$ aussi) avec $v = w$ (théorème 1, $N \neq 37$ encore), donc $u\infty = w0 = \infty$. Donc $u = 1$ si $m = 1$; de même, $u = w$ si $m = -1$.

Un théorème de MAZUR [3] dit qu'il n'y a que cinq possibilités pour m , c'est-à-dire $m = \pm 1$, 0 , $\pm 1/3$; dans notre cas, il est clair que $m \neq 0$. De plus, $m = \pm 1/3$ n'est possible que si $N \equiv -1 \pmod 3$ et $N = 53$, 113 , 137 , donc $n \geq 13$. Mais $u^2 = 1$ implique $m^2 = 1$ dans $\underset{\sim}{Z}/n\underset{\sim}{Z}$, i. e. $1 \equiv 9 \pmod n$, évidemment impossible pour $n \geq 13$. Donc $m = \pm 1$, $u = 1$, w , et on a le théorème 2.

3. Réduction modulo p de $X_0(p)$.

Si E est une courbe elliptique en caractéristique $p > 0$, l'application $p : E \longrightarrow E$ est inséparable de degré p^2 . Il y a donc deux cas :

1° p est totalement inséparable, i. e. $E_p = 0$. Dans ce cas, on dit que E est supersingulière ; son invariant $j = j(E)$ est nécessairement dans $\underset{\sim}{F}_{p^2}$.

2° Le degré de séparabilité de p est p , i. e. E_p est d'ordre p .

En tous cas, E admet l'isogénie de Frobenius $\varphi : E \longrightarrow E^{(p)}$, qui est inséparable de degré p (Si E est définie par une équation $F(x , y) = \sum a_{ij} x^i y^j = 0$,

alors $E^{(p)}$ est définie par $\sum a_{ij}^p x^i y^j = 0$, et $\varphi(x , y) = (x^p , y^p)$).

A un isomorphisme près, φ est l'unique isogénie sur E qui est inséparable de degré p . On note par $\hat{\varphi} = w(\varphi)$ sa transposée $\hat{\varphi} : E^{(p)} \longrightarrow E$ (Donc $\hat{\varphi} \circ \varphi$ est l'endomorphisme p de E). Si E n'est pas supersingulière, alors $\hat{\varphi}$ est l'unique (à un isomorphisme près) isogénie sur $E^{(p)}$ qui est séparable de degré p (Son noyau est $E_p^{(p)}$).

La réduction modulo p de $X_0(p)$ est la réunion de deux droites projectives Z et Z' , se coupant transversalement. Les points de Z sont représentés par des Frobenius $\varphi : E \longrightarrow E^{(p)}$ et les points de Z' par les transposées

$$\hat{\varphi} : E^{(p)} \longrightarrow E .$$

Les intersections de Z et Z' sont les points supersinguliers, où $\hat{\varphi}$ est inséparable. L'involution w échange Z et Z' , et opère sur $Z \cap Z'$ comme le Frobenius. [Voir DELIGNE-RAPOPORT [1] pour les détails. En caractéristique 0 , il est classique que $0 = F(j \circ w , j)$, où $j \circ w(\tau) = j(-1/p\tau) = j(p\tau)$, et $F(X , Y) \in \underset{\sim}{Z}[X , Y]$ est irréductible, et $X_0(p)$ est la normalisée de $F = 0$. On a $F(X , Y) \equiv (X - Y^p)(X^p - Y) \pmod p$, la "congruence de Kronecker-Weber", d'où le fait que les intersections sont transversales ; c'est la démonstration de DELIGNE du fait que le polynôme de Hasse (qui a comme racines les valeurs supersingulières de j) a des racines distinctes.] Notons que $Z + Z'$ n'est pas en général le modèle de Néron, mais suffit pour nos besoins.

Rappelons que le genre arithmétique $p_a(C)$ d'une courbe réductible C satisfait aux règles suivantes :

1° $p_a(C + D) = -1 + p_a(C) + p_a(D) + C.D$, où $C.D$ est le nombre des intersections de C et D (multiplicités comptées).

2° $p_a(C)$ est invariant par spécialisation et par éclatement.

3° Si C est irréductible, et si g est le genre de sa normalisée, alors $p_a(C) \geq g$, avec $p_a(C) = g$ si, et seulement si, C est non-singulière. En particulier, $p_a(C) = 0$ si, et seulement si, $C = \underset{\sim}{P}^1$.

Dans notre cas, $Z + Z'$ est une spécialisation (réduction modulo p) de $X_0(p)$, de genre g en caractéristique 0 . Le 1° donne :

$$1 + g = 1 + p_a(Z + Z') = p_a(Z) + p_a(Z') + Z.Z' ,$$

i. e.

(14) $\quad 1 + g = Z.Z' =$ nombre des valeurs supersingulières de j .

Ecrivons $1 + g = r + 2s$, où r valeurs supersingulières $a_1 , \ldots , a_r$ de j sont dans $\underset{\sim}{F}_p$, et $2s$ valeurs supersingulières $b_1 , b_1^p , \ldots , b_s , b_s^p$ de j dans $\underset{\sim}{F}_{p^2} - \underset{\sim}{F}_p$. L'automorphisme w de $Z + Z'$ échange Z et Z' , fixe $a_1 , \ldots , a_r$, et échange b_i et b_i^p ; le quotient $(Z + Z')/(w)$ est donc une droite C avec s points doubles ordinaires, et donc avec genre arithmétique s ,

comme on le vérifie en éclatant ces points doubles, grâce aux 1° et 2°. Encore par le principe de spécialisation, ce genre arithmétique est g^+ , le genre de $X_0^+(p)$ en caractéristique 0 . D'où :

(15) $2g^+$ = nombre des valeurs supersingulières de j dans $\underset{\sim}{F}_{p^2} - \underset{\sim}{F}_p$.

J'ignore les origines exactes de (15) ; la formule était mentionnée dans le cours de SERRE de 1972/73 au Collège de France, dans le cadre de la théorie des formes modulaires modulo p . En tous cas, on a comme corollaire de (15) et du théorème 1, le résultat suivant.

COROLLAIRE. - Toutes les valeurs supersingulières de j sont dans $\underset{\sim}{F}_p$ si, et seulement si, $g^+ = 0$, i. e. $g \leqslant 1$ où $X_0(p)$ est hyperelliptique avec $v = w$, i. e. $p \leqslant 31$ ou $p = 41 , 47 , 59 , 71$.

Autrement dit, toutes les racines du polynôme de Hasse sont dans $\underset{\sim}{F}_p$, si, et seulement si, p est une de ces quinze valeurs. Notons que d'après les formules (2) et (14), la formule (15) peut être exprimée aussi sous la forme

(16) $n(w)/2$ = nombre des valeurs supersingulières de j dans $\underset{\sim}{F}_p$,

où $n(w)$ est donné par (3).

Remarque 1. - Dans sa leçon d'ouverture au Collège de France, le 14 janvier 1975, J. TITS mentionna le groupe de Fischer, "le monstre", qui, s'il existe, est un groupe simple "sporadique" d'ordre

$$2^{46}.3^{20}.5^9.7^6.11^2.13^3.17.19.23.29.31.41.47.59.71 ,$$

i. e. divisible exactement par les quinze nombres premiers de la liste du corollaire. Une bouteille de Jack Daniels est offerte à celui qui expliquera cette coïncidence.

Remarque 2. - Une des diverses astuces utilisées dans la démonstration du théorème 1 est un résultat de Schoeneberg [9], qui montre que, dans le mauvais cas, où $X_0(N)$ est hyperelliptique avec $v \neq w$, on a $n(w) \leqslant 4$ (cf. (3)). Il est amusant que, dans le cas $N = p$, on puisse donner une démonstration très éloignée de la démonstration originale de Schoeneberg (exercice sur les points de Weierstrass). En effet, dans ce cas, on a $(Z + Z')/(v) = \underset{\sim}{P}^1$, i. e. v échange Z et Z' , et fixe tous les points d'intersection $a_1 , \ldots , a_r , b_1 , b_1^p , \ldots , b_s , b_s^p$. Alors l'involution $u = v.w$ fixe Z et Z' et $a_1 , \ldots , a_r$. Un automorphisme non trivial de $\underset{\sim}{P}^1$ ayant $\leqslant 2$ points fixes, on a $r \leqslant 2$, i. e. $n(w) \leqslant 4$, d'après (16).

BIBLIOGRAPHIE

[1] DELIGNE (P.) et RAPOPORT (M.). - Les schémas de modules de courbes elliptiques, "Modular functions of one variable, II", p. 143-316. - Berlin, Heidelberg, New York, Springer-Verlag, 1973 (Lecture Notes in Mathematics, 349).

[2] LEHNER (J.) and NEWMAN (M.). - Weierstrass points of $\Gamma_0(N)$, Annals of Math., t. 79, 1964, p. 360-368.

[3] MAZUR (B). - Modular curves and the Eisenstein ideal (en préparation).

[4] MAZUR (B.). - Lettre à Andrew Ogg (datée du 25 novembre 1974).

[5] MAZUR (B.) and SWINNERTON-DYER (P.). - Arithmetic of Weil curves, Invent. Math., Berlin, t. 25, 1974, p. 1-61.

[6] OGG (A.). - Hyperelliptic modular curves, Bull. Soc. math. France, t. 102, 1974, p. 449-462.

[7] OGG (A.). - Diophantine equations and modular forms, Bull. Amer. math. Soc., t. 81, 1975, p. 14-27.

[8] RIBET (K.). - Lettre à Andrew Ogg (datée du 29 novembre 1974, reçue le 27 décembre 1974).

[9] SCHOENEBERG (B.). - Über die Weierstrasspunkte in den Körpern der elliptischen Modulfunktionen, Abh. Math. Semin. Hamburg, t. 17, 1951, p. 104-111.

(Texte reçu le 30 janvier 1975)

Andrew P. OGG
Department of Mathematics
University of California
BERKELEY, Calif. 94720
(Etats-Unis)

BTZ black hole

The **BTZ black hole**, named after Máximo Bañados, Claudio Teitelboim, and Jorge Zanelli, is a black hole solution for (2+1)-dimensional topological gravity with a negative cosmological constant.

Contents

History

Properties

The case without charge

See also

References

History

In 1992 Bañados, Teitelboim and Zanelli discovered the BTZ black hole solution (Bañados, Teitelboim & Zanelli 1992). This came as a surprise, because when the cosmological constant is zero, a vacuum solution of (2+1)-dimensional gravity is necessarily flat (the Weyl tensor vanishes in three dimensions, while the Ricci tensor vanishes due to the Einstein field equations, so the full Riemann tensor vanishes), and it can be shown that no black hole solutions with event horizons exist. But thanks to the negative cosmological constant in the BTZ black hole, it is able to have remarkably similar properties to the 3+1 dimensional Schwarzschild and Kerr black hole solutions, which model real-world black holes.

Properties

The similarities to the ordinary black holes in 3+1 dimensions:

- It admits a no hair theorem, fully characterizing the solution by its ADM-mass, angular momentum and charge.
- It has the same thermodynamical properties as traditional black hole solutions such as Schwarzschild or Kerr black holes, e.g. its entropy is captured by a law directly analogous to the Bekenstein bound in (3+1)-dimensions, essentially with the surface area replaced by the BTZ black hole's circumference.
- Like the Kerr black hole, a rotating BTZ black hole contains an inner and an outer horizon, analogous to an ergosphere.

Since (2+1)-dimensional gravity has no Newtonian limit, one might fear that the BTZ black hole is not the final state of a gravitational collapse. It was however shown, that this black hole could arise from collapsing matter and we can calculate the energy-moment tensor of BTZ as same as (3+1) black holes. (Carlip 1995) section 3 Black Holes and Gravitational Collapse.

The BTZ solution is often discussed in the realm on (2+1)-dimensional quantum gravity.

The case without charge

The metric in the absence of charge is

$$ds^2 = -\frac{(r^2 - r_+^2)(r^2 - r_-^2)}{l^2 r^2} dt^2 + \frac{l^2 r^2 dr^2}{(r^2 - r_+^2)(r^2 - r_-^2)} + r^2 \left(d\phi - \frac{r_+ r_-}{l r^2} dt \right)^2$$

where r_+, r_- are the black hole radii and l is the radius of AdS_3 space. The mass and angular momentum of the black hole is

$$M = \frac{r_+^2 + r_-^2}{l^2}, \quad J = \frac{2 r_+ r_-}{l}$$

BTZ black holes without any electric charge are locally isometric to anti-de Sitter space. More precisely, it corresponds to an orbifold of the universal covering space of AdS_3.

A rotating BTZ black hole admits closed timelike curves.

See also

- Cosmic string
- MTZ black hole
- AdS black hole

References

Notes

Bibliography

- Bañados, Máximo; Teitelboim, Claudio; Zanelli, Jorge (28 September 1992), "The Black hole in three-dimensional space-time", *Phys. Rev. Lett.*, **69** (13): 1849–51, arXiv:hep-th/9204099v3 (https://arxiv.org/abs/hep-th/9204099v3), Bibcode:1992PhRvL..69.1849B (https://ui.adsabs.harvard.edu/abs/1992PhRvL..69.1849B), doi:10.1103/PhysRevLett.69.1849 (https://doi.org/10.1103%2FPhysRevLett.69.1849), PMID 10046331 (https://pubmed.ncbi.nlm.nih.gov/10046331), S2CID 18095488 (https://api.semanticscholar.org/CorpusID:18095488)
- Carlip, Steven (2005), "Conformal Field Theory, (2+1)-Dimensional Gravity, and the BTZ Black Hole", *Classical and Quantum Gravity*, **22** (12): R85–R123, arXiv:gr-qc/0503022v4 (https://arxiv.org/abs/gr-qc/0503022v4), Bibcode:2005CQGra..22R..85C (https://ui.adsabs.harvard.edu/abs/2005CQGra..22R..85C), doi:10.1088/0264-9381/22/12/R01 (https://doi.org/10.1088%2F0264-9381%2F22%2F12%2FR01), S2CID 115762178 (https://api.semanticscholar.org/CorpusID:115762178)
- Carlip, Steven (1995), "The (2+1)-Dimensional Black Hole", *Classical and Quantum Gravity*, **12** (12): 2853–2879, arXiv:gr-qc/9506079 (https://arxiv.org/abs/gr-qc/9506079), Bibcode:1995CQGra..12.2853C (https://ui.adsabs.harvard.edu/abs/1995CQGra..12.2853C),

doi:10.1088/0264-9381/12/12/005 (https://doi.org/10.1088%2F0264-9381%2F12%2F12%2F005), S2CID 119508585 (https://api.semanticscholar.org/CorpusID:119508585)

- Bañados, Máximo (1999), "Three-dimensional quantum geometry and black holes" (http://cds.cern.ch/record/377704/files/9901148.pdf) (PDF), *Trends in Theoretical Physics II*, AIP Conference Proceedings, **484**: 147–169, arXiv:hep-th/9901148v3 (https://arxiv.org/abs/hep-th/9901148v3), Bibcode:1999AIPC..484..147B (https://ui.adsabs.harvard.edu/abs/1999AIPC..484..147B), doi:10.1063/1.59661 (https://doi.org/10.1063%2F1.59661), S2CID 7598959 (https://api.semanticscholar.org/CorpusID:7598959)
- Ida, Daisuke (30 October 2000), "No Black Hole Theorem in Three-Dimensional Gravity", *Phys. Rev. Lett.*, **85** (18): 3758–60, arXiv:gr-qc/0005129 (https://arxiv.org/abs/gr-qc/0005129), Bibcode:2000PhRvL..85.3758I (https://ui.adsabs.harvard.edu/abs/2000PhRvL..85.3758I), doi:10.1103/PhysRevLett.85.3758 (https://doi.org/10.1103%2FPhysRevLett.85.3758), PMID 11041920 (https://pubmed.ncbi.nlm.nih.gov/11041920), S2CID 38770795 (https://api.semanticscholar.org/CorpusID:38770795)

Retrieved from "https://en.wikipedia.org/w/index.php?title=BTZ_black_hole&oldid=1107063403"

This page was last edited on 27 August 2022, at 23:28 (UTC).

Text is available under the Creative Commons Attribution-ShareAlike License 3.0; additional terms may apply. By using this site, you agree to the Terms of Use and Privacy Policy. Wikipedia® is a registered trademark of the Wikimedia Foundation, Inc., a non-profit organization.

hep-th/yymmnnn

Three-Dimensional Gravity Reconsidered

arXiv:0706.3359v1 [hep-th] 22 Jun 2007

Edward Witten

School of Natural Sciences, Institute for Advanced Study
Princeton, New Jersey 08540

We consider the problem of identifying the CFT's that may be dual to pure gravity in three dimensions with negative cosmological constant. The c-theorem indicates that three-dimensional pure gravity is consistent only at certain values of the coupling constant, and the relation to Chern-Simons gauge theory hints that these may be the values at which the dual CFT can be holomorphically factorized. If so, and one takes at face value the minimum mass of a BTZ black hole, then the energy spectrum of three-dimensional gravity with negative cosmological constant can be determined exactly. At the most negative possible value of the cosmological constant, the dual CFT is very likely the monster theory of Frenkel, Lepowsky, and Meurman. The monster theory may be the first in a discrete series of CFT's that are dual to three-dimensional gravity. The partition function of the second theory in the sequence can be determined on a hyperelliptic Riemann surface of any genus. We also make a similar analysis of supergravity.

June, 2007

Contents

1. Introduction . . . 1
   1.1. Relation To Gauge Theory . . . 2
   1.2. What To Aim For . . . 5
   1.3. A Non-Classical Restriction . . . 6
   1.4. Plan Of This Paper . . . 7
2. Gauge Theory And The Value Of c . . . 9
   2.1. Quantization Of Parameters . . . 11
   2.2. Comparison To Three-Dimensional Gravity . . . 15
   2.3. Holomorphic Factorization . . . 16
   2.4. Interpretation . . . 19
   2.5. Analog For Supergravity . . . 25
3. Partition Functions . . . 28
   3.1. The Bosonic Case . . . 29
   3.2. The Supersymmetric Case . . . 38
   3.3. Extremal SCFT's With Small k^* . . . 47
4. $k = 2$ Partition Function On A Hyperelliptic Riemann Surface . . . 58
   4.1. Twist Fields . . . 59
   4.2. The $\mathcal{E} \cdot \mathcal{E}$ Operator Product . . . 63
   4.3. Determining The Four-Point Function . . . 66
   4.4. The Genus One Partition Function Revisited . . . 70
   4.5. Role Of The Conformal Anomaly . . . 72
Appendix A. More Extremal Partition Functions . . . 76

1. Introduction

Three-dimensional pure quantum gravity, with the Einstein-Hilbert action

$$I = \frac{1}{16\pi G} \int d^3x \sqrt{g} \left(R + \frac{2}{\ell^2} \right), \tag{1.1}$$

has been studied from many points of view (see [1-3] for some early developments and [4] for a recent review and references), but its status is fundamentally unclear. The present paper is devoted to a tentative attempt to reconsider it. Before giving an overview of the paper, we begin with an introduction to the problem.

The first thought about this theory is that at the classical level it is "trivial," in the sense that there are no gravitational waves, and any two solutions are equivalent locally. So perhaps it might be tractable quantum mechanically.

A second thought is that despite being "trivial," the theory actually is unrenormalizable by power counting, since the gravitational constant G has dimensions of length. So perhaps the quantum theory does not exist.

The claim about unrenormalizability, however, is fallacious, precisely because the classical theory is trivial. In three dimensions, the Riemann tensor R_{ijkl} can be expressed in terms of the Ricci tensor R_{ij}. In the case of pure gravity, the equations of motion express the Ricci tensor as a constant times the metric. So finally, any possible counterterm can be reduced to a multiple of $\int d^3x\sqrt{g}$ and is equivalent on-shell to a renormalization of the cosmological constant, which is parametrized in (1.1) via the parameter ℓ^2. A counterterm that vanishes on shell can be removed by a local redefinition of the metric tensor g (of the general form $g_{ij} \to g_{ij} + aR_{ij} + \ldots$, where a is a constant and the ellipses refer to local terms of higher order). So a more precise statement is that any divergences in perturbation theory can be removed by a field redefinition and a renormalization of ℓ^2.

1.1. Relation To Gauge Theory

The claim just made is valid regardless of how one formulates perturbation theory. But actually, there is a natural formulation in which no field redefinition or renormalization is needed. This comes from the fact that classically, $2+1$-dimensional pure gravity can be expressed in terms of gauge theory. The spin connection ω is an $SO(2,1)$ gauge field (or an $SO(3)$ gauge field in the case of Euclidean signature). It can be combined with the "vierbein" e to make a gauge field of the group $SO(2,2)$ if the cosmological constant is negative (and a similar group if the cosmological constant is zero or positive). We simply combine ω and e to a 4×4 matrix A of one-forms:

$$A = \begin{pmatrix} \omega & e/\ell \\ -e/\ell & 0 \end{pmatrix}. \tag{1.2}$$

Here ω fills out a 3×3 block, while e occupies the last row and column. As long as e is invertible, the usual transformations of e and ω under infinitesimal local Lorentz transformations and diffeomorphisms combine together into gauge transformations of A. This statement actually has a close analog in any spacetime dimension d, with $SO(d-1,2)$ replacing $SO(2,2)$. What is special in $d=3$ is that [5,6] it is also possible to write the action in a gauge-invariant form. Indeed the usual Einstein-Hilbert action (1.1) is equivalent to a Chern-Simons Lagrangian[1] for the gauge field A:

$$I = \frac{k}{4\pi}\int \mathrm{tr}^*\left(A\wedge dA + \frac{2}{3}A\wedge A\wedge A\right). \tag{1.3}$$

[1] Here tr^* denotes an invariant quadratic form on the Lie algebra of $SO(2,2)$, defined by $\mathrm{tr}^*\, ab = \mathrm{tr}\; a\star b$, where tr is the trace in the four-dimensional representation and $\star$ is the Hodge star, $(\star b)_{ij} = \frac{1}{2}\epsilon_{ijkl}b^{kl}$.

In dimensions other than three, it is not possible to similarly replace the Einstein-Hilbert action with a gauge-invariant action for gauge fields.[2]

In the gauge theory description, perturbation theory is renormalizable by power counting, and is actually finite, because there are no possible local counterterms. The Chern-Simons functional itself is the only gauge-invariant action that can be written in terms of A alone without a metric tensor; as it is not the integral of a gauge-invariant local density, it will not appear as a counterterm in perturbation theory. The cosmological constant cannot be renormalized, since in the gauge theory description, it is a structure constant of the gauge group.

As we have already remarked, the Chern-Simons description of three-dimensional gravity is valid when the vierbein is invertible. This is so for a classical solution, so it is true if one is sufficiently close to a classical solution. Perturbation theory, starting with a classical solution, will not take us out of the region in which the vierbein is invertible, so the Chern-Simons description of three-dimensional gravity is valid perturbatively. The fact that, in this formulation, the perturbation expansion of three-dimensional gravity is actually finite can reasonably be taken as a hint that the corresponding quantum theory really does exist.

However, nonperturbatively, the relation between three-dimensional gravity and Chern-Simons gauge theory is unclear. For one thing, in Chern-Simons theory, nonperturbatively the vierbein may cease to be invertible. For example, there is a classical solution with $A = \omega = e = 0$. The viewpoint in [6] was that such non-geometrical configurations must be included to make sense of three-dimensional quantum gravity nonperturbatively. But it has has been pointed out (notably by N. Seiberg) that when we do know how to make sense of quantum gravity, we take the invertibility of the vierbein seriously. For example, in perturbative string theory, understood as a model of quantum gravity in two spacetime dimensions, the integration over moduli space of Riemann surfaces that leads to a sensible theory is derived assuming that the metric should be non-degenerate.

There are other possible problems in the nonperturbative relation between three-dimensional gravity and Chern-Simons theory. The equivalence between diffeomorphisms and gauge transformations is limited to diffeomorphisms that are continuously connected to the identity. However, in gravity, we believe that more general diffeomorphisms (such as

[2] In $d = 4$, it is possible to write the Hamiltonian constraints of General Relativity in terms of gauge fields [7]. This has been taken as the starting point of loop quantum gravity.

modular transformations in perturbative string theory) play an important role. These are not naturally incorporated in the Chern-Simons description. One can by hand supplement the gauge theory description by imposing invariance under disconnected diffeomorphisms, but it is not clear how natural this is.

Similarly, in quantum gravity, one expects that it is necessary to sum over the different topologies of spacetime. Nothing in the Chern-Simons description requires us to make such a sum. We can supplement the Chern-Simons action with an instruction to sum over three-manifolds, but it is not clear why we should do this.

From the point of view of the Chern-Simons description, it seems natural to fix a particular Riemann surface Σ, say of genus g, and construct a quantum Hilbert space by quantizing the Chern-Simons gauge fields on Σ. (Indeed, there has been remarkable progress in learning how to do this and to relate the results to Liouville theory [8-11].) In quantum gravity, we expect topology-changing processes, such that it might not be possible to associate a Hilbert space with a particular spatial manifold.

Regardless of one's opinion of questions such as these, there is a more serious problem with the idea that gravity and gauge theory are equivalent non-perturbatively in three dimensions. Some years after the gauge/gravity relation was suggested, it was discovered by Bañados, Teitelboim, and Zanelli [12] that in three-dimensional gravity with negative cosmological constant, there are black hole solutions. The existence of these objects, generally called BTZ black holes, is surprising given that the classical theory is "trivial." Subsequent work [13,14] has made it clear that three-dimensional black holes should be taken seriously, particularly in the context of the AdS/CFT correspondence [15].

The BTZ black hole has a horizon of positive length and a corresponding Bekenstein-Hawking entropy. If, therefore, three-dimensional gravity does correspond to a quantum theory, this theory should have a huge degeneracy of black hole states. It seems unlikely that this degeneracy can be understood in Chern-Simons gauge theory, because this essentially topological theory has too few degrees of freedom. However, some interesting attempts have been made; for a review, see [4].

The existence of the BTZ black hole makes three-dimensional gravity a much more exciting problem. This might be our best chance for a solvable model with quantum black holes. Surely in 3+1 dimensions, the existence of gravitational waves with their nonlinear interactions means that one cannot hope for an exact solution of any system that includes

quantum gravity.[3] There might be an exact solution of a 1+1-dimensional model with black holes (interesting attempts have been made [16]), but such a model is likely to be much less realistic than three-dimensional pure gravity. For example, in 1+1 dimensions, the horizon of a black hole just consists of two points, so there is no good analog of the area of the black hole horizon.

1.2. What To Aim For

So we would like to solve three-dimensional pure quantum gravity. But what would it mean to solve it?

First of all, we will only consider the case that the cosmological constant Λ is negative. This is the only case in which we know what it would mean to solve the theory.

Currently, there is some suspicion (for example, see [17]) that quantum gravity with $\Lambda > 0$ does not exist nonperturbatively, in any dimension. One reason is that it does not appear possible with $\Lambda > 0$ to define precise observables, at least none [18] that can be measured by an observer in the spacetime.[4] This is natural if a world with positive Λ (like the one we may be living in) is always at best metastable – as is indeed the case for known embeddings of de Sitter space in string theory [20]. If that is so, then pure gravity with $\Lambda > 0$ does not really make sense as an exact theory in its own right but (like an unstable particle) must be studied as part of a larger system. There may be many choices of the larger system (for example, many embeddings in string theory) and it may be unrealistic to expect any of them to be soluble.

Whether that is the right interpretation or not, we cannot in this paper attempt to solve three-dimensional gravity with $\Lambda > 0$, since, not knowing how to define any mathematically precise observables, we do not know what to try to calculate.

For $\Lambda = 0$, above three dimensions there is a precise observable in quantum gravity, the S-matrix. However, in the three-dimensional case, there is no S-matrix in the usual sense, since in any state with nonzero energy, the spacetime is only locally asymptotic to Minkowski space at infinity [3]. More relevantly for our purposes, in three-dimensional pure gravity with $\Lambda = 0$, there is no S-matrix since there are no particles that can be

[3] An exact solution or at least an illuminating description of the appropriate Hamiltonian for near-extremal black holes interacting with external massless particles is conceivable.

[4] At least perturbatively, the de Sitter/CFT correspondence gives observables that can be measured by an observer who looks at the whole universe from the outside [19,18]. These observables characterize the wavefunction of the ground state.

scattered. There are no gravitons in three dimensions, and there are also no black holes unless $\Lambda < 0$. So again, we do not have a clear picture of what we would aim for to solve three-dimensional gravity with zero cosmological constant.

With negative cosmological constant, there is an analog, and in fact a much richer analog, of the S-matrix, namely the dual conformal field theory (CFT). This is of course a *two*-dimensional CFT, defined on the asymptotic boundary of spacetime. Not only does AdS/CFT duality make sense in three dimensions, but in fact one of the precursors of the AdS/CFT correspondence was the discovery by Brown and Henneaux [21] of an asymptotic Virasoro algebra in three-dimensional gravity. They considered three-dimensional gravity with negative cosmological constant possibly coupled to additional fields. The action is

$$I = \frac{1}{16\pi G}\int d^3x\sqrt{g}\left(R + \frac{2}{\ell^2} + \dots\right), \tag{1.4}$$

where the ellipses reflect the contributions of other fields. Their main result is that the physical Hilbert space obtained in quantizing this theory (in an asymptotically Anti de Sitter or AdS spacetime) has an action of left- and right-moving Virasoro algebras with $c_L = c_R = 3\ell/2G$. In our modern understanding [15], this is part of a much richer structure – the boundary conformal field theory.

What it means to solve pure quantum gravity with $\Lambda < 0$ is to find this dual conformal field theory. We focus on the case $\Lambda < 0$ because this is the only case in which we would know what it means to solve the theory. Luckily, and perhaps not coincidentally, this is also the case that has black holes.

1.3. A Non-Classical Restriction

This formulation of what we aim to do makes it clear that we must anticipate a restriction that is rather surprising from a classical point of view. In contemplating the classical action (1.4), it appears that the dimensionless ratio ℓ/G is a free parameter. But the formula for the central charge $c_L = c_R = 3\ell/2G$ shows that this cannot be the case. According to the Zamolodchikov c-theorem [22], in any continuously varying family of conformal field theories in $1+1$ dimensions, the central charge c is constant. More generally, the same is true for the left- and right-moving central charges c_L and c_R.

So the central charges of the dual CFT cannot depend on a continuously variable parameter ℓ/G. It must be [18] that the theory only makes sense for specific values of ℓ/G.

Of course, the c-theorem has an important technical assumption: the theory must have a normalizable and $SL(2,\mathbb{R}) \times SL(2,\mathbb{R})$-invariant ground state. (The two factors of $SL(2,\mathbb{R})$ are for left- and right-moving boundary excitations.) This condition is obeyed by three-dimensional gravity, with Anti de Sitter space being the classical approximation to the vacuum. The desired $SL(2,\mathbb{R}) \times SL(2,\mathbb{R})$ symmetry is simply the classical $SO(2,2)$ symmetry of three-dimensional AdS space.

The statement that ℓ/G cannot be continuously varied is not limited to pure gravity. It holds for the same reason in any theory of three-dimensional gravity plus matter that has a sensible AdS vacuum. For example, in the string theory models whose CFT duals are known, ℓ/G is expressed in terms of integer-valued fluxes; this gives a direct explanation of why it cannot be varied continuously.

1.4. Plan Of This Paper

Now we can describe the plan of this paper.

We aim to solve three-dimensional gravity with negative Λ, at some distinguished values of ℓ/G at which it makes sense.

We do not have any rigorous way to determine the right values. However, in section 2, taking at face value the Chern-Simons description of three-dimensional gravity, we will use it to motivate certain values of ℓ/G. The values that emerge – with the help of a small sleight of hand in the choice of the gauge group for the Chern-Simons theory – are interesting. They are the values at which c_L and c_R are integer multiples of 24, and complete holomorphic factorization of the dual CFT is conceivable.

Not wishing to look a gift horse in the mouth, we will assume that these are the right values to consider. Relying on holomorphic factorization, to describe the solution of the theory, we must describe a sequence of holomorphic CFT's with $c = 24k$, $k = 1, 2, 3, \ldots$.

For $k = 1$, it is believed [23] that there are precisely 71 holomorphic CFT's with the relevant central charge $c = 24$. Of these theories, 70 have some form of Kac-Moody or current algebra symmetry extending the conformal symmetry. In the AdS/CFT correspondence, these theories are dual to three-dimensional theories describing gravity plus additional gauge fields (with Chern-Simons interactions). To describe pure gravity, we need a holomorphic CFT with $c = 24$ and no Kac-Moody symmetry.

Such a model was constructed nearly twenty-five years ago by Frenkel, Lepowsky, and Meurman [24], who also conjectured its uniqueness. The motivation for constructing the model was that it admits as a group of symmetries the Fischer-Griess monster group $\mathbb{M}$ –

the largest of the sporadic finite groups. (The link between the monster and conformal field theory was suggested by developments springing from an observation by McKay relating the monster to the j-function, as we explain more fully in section 3.1.) Arguably, the FLM model is the most natural known structure with $\mathbb{M}$ symmetry. A detailed and elegant description is in the book [25]; for short summaries, see [26,27], and for subsequent developments and surveys, see [28-30]. Assuming the (unproved) uniqueness conjecture of FLM, we propose that their model must give the CFT that is dual to three-dimensional gravity at $c = 24$.

For $c = 24k$, $k > 1$, we need an analog of requiring that there is no Kac-Moody symmetry. A plausible analog, expressing the idea that we aim to describe pure gravity, is that there should be no primary fields of low dimension other than the identity. A small calculation shows that at $c = 24k$, the lowest dimension of a primary other than the identity cannot be greater than $k + 1$, and if we assume that this dimension is precisely $k + 1$, then the partition function is uniquely determined. Conformal field theories with this property were first investigated by Höhn in [31,32] and have been called extremal CFT's; see also [33].

It is not known[5] if extremal CFT's exist for $k > 1$. If such a CFT does exist, it is an attractive candidate for the dual of three-dimensional gravity at the appropriate value of the cosmological constant. The primaries of dimension $k + 1$ and above would be interpreted as operators that create black holes. The dimension $k + 1$ agrees well with the minimum mass of a BTZ black hole. This statement may sound like magic, since the value $k + 1$ is determined from holomorphy and modular invariance without mentioning black holes; but the result is not so surprising if one is familiar with previous results on the AdS/CFT correspondence in three dimensions [34].

Section 3 of this paper is devoted to describing the partition function of an extremal CFT and discussing how such a theory could be related to three-dimensional gravity. In both sections 2 and 3, we consider also the case of three-dimensional supergravity. More precisely, we consider only minimal supergravity, corresponding to $N = 1$ superconformal symmetry for the boundary CFT. In this case, holomorphic factorization is conceivable at $c = 12k^*$, $k^* = 1, 2, 3, \ldots$. Here there is a little ambiguity about exactly what we

[5] Höhn's definition of an extremal CFT allowed holomorphic factorization up to a phase, so that c may be a multiple of 8, not 24. As a result, he discussed several examples of extremal theories that we will not consider here. These examples have k non-integral and less than 2.

should mean by an extremal superconformal field theory (SCFT), but pragmatically, there are good candidates at $k^* = 1, 2$. The $k^* = 1$ theory was constructed by Frenkel, Lepowsky, and Meurman, who also conjectured its uniqueness. Its discrete symmetries were understood only recently in work by Duncan [35]. For $k^* = 2$, the extremal SCFT was constructed by Dixon, Ginsparg, and Harvey [26], by modifying the orbifold projection that had been used [24] in constructing the $k = 1$ extremal CFT. Interestingly, the extremal SCFT's with $k^* = 1, 2$ both admit[6] an action of very large discrete groups related to the Conway group. This is a further indication that unusual discrete groups are relevant to three-dimensional gravity and supergravity. In fact, we find some hints that supergravity may have monster symmetry at $k^* = 4$ and baby monster symmetry at $k^* = 6$.

Regrettably, we do not know how to construct new examples of extremal conformal or superconformal field theories. Section 4 is devoted to a calculation that aims to give modest support to the idea that new extremal theories do exist. We consider an extremal CFT with $k = 2$ and show that its partition function can be uniquely determined on a hyperelliptic Riemann surface of any genus (including, for example, any Riemann surface of genus 2). The fact that a partition function with the right properties exists and is unique for any genus is hopefully a hint that an extremal $k = 2$ CFT does exist.

We make at each stage the most optimistic possible assumption. Decisive arguments in favor of the proposals made here are still lacking. The literature on three-dimensional gravity is filled with claims (including some by the present author [6]) that in hindsight seem less than fully satisfactory. Hopefully, future work will clarify things.

Advice by J. Maldacena was essential at the outset of this work. I also wish to thank J. Duncan, G. Höhn, G. Nebe, and J. Teschner for descriptions of their work and helpful advice; T. Gannon, R. Griess, J. Lepowsky, and A. Ryba for correspondence about the monster group and related matters; and many colleagues at the IAS and elsewhere, especially A. Maloney, G. Moore, and S. Shenker, for helpful comments.

2. Gauge Theory And The Value Of c

The goal of the present section is to determine what values of the cosmological constant, or equivalently of the central charge c of the boundary CFT, are suggested by the relation between three-dimensional gravity and Chern-Simons gauge theory.

[6] The fact that they have essentially the same symmetry group was pointed out by J. Duncan, who also suggested the identification of the $k^* = 2$ theory.

Before proceeding to any calculation, we will dispose of a few preliminary points. The first is that [36], as long as the three-dimensional spacetime is oriented, as we will assume in this paper, three-dimensional gravity can be generalized to include an additional interaction, the Chern-Simons functional of the spin connection ω:

$$\Delta_0 I = \frac{k'}{4\pi}\int_W \mathrm{tr}\,\left(\omega\wedge d\omega+\frac{2}{3}\omega\wedge\omega\wedge\omega\right). \tag{2.1}$$

Here we think of ω as an $SO(2,1)$ gauge field (or an $SO(3)$ gauge field in the case of Euclidean signature). Also, tr is the trace in the three-dimensional representation of $SO(2,1)$, and k' is quantized for topological reasons (the precise normalization depends on some assumptions and is discussed in sections 2.1 and 2.4). Equivalently, instead of ω, we could use the $SO(2,2)$ gauge field A introduced in eqn. (1.2), and add to the action a term of the form

$$\Delta I = \frac{k'}{4\pi}\int_M \mathrm{tr}\,\left(A\wedge dA+\frac{2}{3}A\wedge A\wedge A\right), \tag{2.2}$$

where now tr is the trace in the four-dimensional representation of $SO(2,2)$. Provided that the conventional Einstein action (1.1) is also present, it does not matter which form of the gravitational Chern-Simons interaction we use, since they lead to equivalent theories. If one adds[7] to ω a multiple of e, the Einstein action (1.1) transforms in a way that cancels the e-dependent part of (2.2), reducing it to (2.1) (while modifying the parameters in the Einstein action).

For our purposes, the $SO(2,2)$-invariant form (2.2) is more useful. This way of writing the Chern-Simons functional places it precisely in parallel with the Einstein-Hilbert action, which as in (1.3) can similarly be expressed as a Chern-Simons interaction, defined with a different quadratic form. We will use the fact that all interactions can be written as Chern-Simons interactions to constrain the proper quantization of all dimensionless parameters, including ℓ/G.

We start with the fact that the group $SO(2,2)$ is locally equivalent to $SO(2,1)\times SO(2,1)$. Moreover, we will in performing the computation assume to start with that $SO(2,1)\times SO(2,1)$ is the right global form of the gauge group. (Then we consider covering

[7] This operation is invariant under diffeomorphisms and local Lorentz transformations because ω and e transform in the same way under local Lorentz transformations – a statement that holds precisely in three spacetime dimensions.

groups[8] that are only locally isomorphic to $SO(2,1) \times SO(2,1)$.) Thus, by taking suitable linear combinations of ω and e, we will obtain a pair of $SO(2,1)$ gauge fields A_L and A_R. These have Chern-Simons interactions

$$I = \frac{k_L}{4\pi} \int \mathrm{tr} \left(A_L \wedge dA_L + \frac{2}{3} A_L \wedge A_L \wedge A_L \right) - \frac{k_R}{4\pi} \int \mathrm{tr} \left(A_R \wedge dA_R + \frac{2}{3} A_R \wedge A_R \wedge A_R \right). \tag{2.3}$$

Both k_L and k_R are integers for topological reasons, and this will lead to a quantization of the ratio G/ℓ that appears in the Einstein-Hilbert action, as well as the gravitational Chern-Simons coupling (2.2). The minus sign multiplying the last term in (2.3) is convenient; it will ensure that k_L and k_R are both positive.

2.1. Quantization Of Parameters

For completeness, we begin by reviewing the quantization of the Chern-Simons coupling in gauge theory. The basic case to consider is that the gauge group is $U(1)$. The gauge field A is a connection on a complex line bundle $\mathcal{L}$ over a three-manifold W, which for simplicity we will assume to have no boundary. Naively speaking, the Chern-Simons action is

$$I = \frac{k}{2\pi} \int_W A \wedge dA \tag{2.4}$$

with some coefficient k. If the line bundle $\mathcal{L}$ is trivial, then we can interpret A as a one-form, and I is well-defined as a real-valued functional. If this were the general situation, there would be no need to quantize k.

However, in general $\mathcal{L}$ is non-trivial, A has Dirac string singularities, and the formula (2.4) is not really well-defined as written. To do better, we pick a four-manifold M of boundary W and such that $\mathcal{L}$ extends over M. Such an M always exists. Then we pick an extension of $\mathcal{L}$ and A over M, and replace the definition (2.4) with

$$I_M = \frac{k}{2\pi} \int_M F \wedge F, \tag{2.5}$$

where $F = dA$ is the curvature. Now there is no Dirac string singularity, and the definition of I_M makes sense. But I_M does depend on M (and on the chosen extension of $\mathcal{L}$, though we do not indicate this in the notation). To quantify the dependence on M, we consider

[8] For an early treatment of coverings in the context of Chern-Simons theory with compact gauge group, see [37]

two different four-manifolds M and M' with boundary W and chosen extensions of $\mathcal{L}$. We can build a four-manifold X with no boundary by gluing together M and M' along W, with opposite orientation for M' so that they fit smoothly along their common boundary. Then we get

$$I_M - I_{M'} = \frac{k}{2\pi}\int_X F\wedge F. \tag{2.6}$$

Now, on the closed four-manifold X, the quantity $\int_X F\wedge F/(2\pi)^2$ represents $\int_X c_1(\mathcal{L})^2$ (here c_1 is the first Chern class) and so is an integer. In quantum mechanics, the action function I should be defined modulo 2π (so that $\exp(iI)$, which appears in the path integral, is single-valued). Requiring $I_M - I_{M'}$ to be an integer multiple of 2π, we learn that k must be an integer. This is the quantization of the Chern-Simons coupling for $U(1)$ gauge theory.[9]

Now let us move on to the case of gauge group $SO(2,1)$. The group $SO(2,1)$ is contractible onto its maximal compact subgroup $SO(2)$, which is isomorphic to $U(1)$. So quantization of the Chern-Simons coupling for an $SO(2,1)$ gauge field can be deduced immediately from the result for $U(1)$. Let A be an $SO(2,1)$ gauge field and define the Chern-Simons coupling

$$I = \frac{k}{4\pi}\int_W \mathrm{tr}\left(A\wedge dA + \frac{2}{3}A\wedge A\wedge A\right), \tag{2.7}$$

where tr is the trace in the three-dimensional representation of $SO(2,1)$. Then, in order for I to be part of the action of a quantum theory, k must be an integer. The reason that the factor $1/2\pi$ in (2.5) has been replaced by $1/4\pi$ in (2.7) is simply that, when we identify $U(1)$ with $SO(2)$ and then embed it in $SO(2,1)$, the trace gives a factor of 2.

Coverings

So we have obtained the appropriate quantization of the Chern-Simons coupling for gauge group $SO(2,1)$. However, this is not quite the whole story, because $SO(2,1)$ is not simply-connected. As it is contractible to $SO(2)\cong U(1)$, it has the same fundamental group as $U(1)$, namely $\mathbb{Z}$. Hence it is possible, for every positive integer n, to take an

[9] There is a refinement if the three-manifold W is endowed with a spin structure. In this case, k can be a half-integer, as explained in [38]. This refinement is physically realized in the quantum Hall effect with filling fraction 1. That effect can be described by an electromagnetic Chern-Simons coupling with $k = 1/2$; the half-integral value is consistent because the microscopic theory has fermions and so requires a spin structure.

n-fold cover of $SO(2,1)$. The most familiar of these is the two-fold cover, $SL(2,\mathbb{R})$. In addition, $SO(2,1)$ has a simply-connected universal cover.

We want to work out the quantization of the Chern-Simons interaction if $SO(2,1)$ is replaced by one of these covering groups. Again, it is convenient to start with $U(1)$. To say that the gauge group of an abelian gauge theory is $U(1)$ rather than $\mathbb{R}$ means precisely that the possible electric charges form a lattice, generated by a fundamental charge that we call "charge 1." Dually, the magnetic fluxes are quantized, with $\int_C F/2\pi \in \mathbb{Z}$ for any two-cycle C. Replacing $U(1)$ by an n-fold cover means that the electric charges take values in $n^{-1}\mathbb{Z}$, and dually, the magnetic fluxes are divisible by n, $\int_C F/2\pi \in n\mathbb{Z}$. As a result, for a four-manifold X, we have $\int_C F\wedge F/(2\pi)^2 \in n^2\mathbb{Z}$. So, in requiring that the Chern-Simons function (2.4) should be well-defined modulo 2π, we require that $k \in n^{-2}\mathbb{Z}$. This is the appropriate result for an n-fold cover. In the case of the universal cover, with $U(1)$ replaced by $\mathbb{R}$, the magnetic fluxes vanish and there is no topological restriction on k.

These statements carry over immediately to covers of $SO(2,1)$, whose covers are all contractible to corresponding covers of $U(1)$. So for an n-fold cover of $SO(2,1)$, we require

$$k \in n^{-2}\mathbb{Z}, \tag{2.8}$$

and for the universal cover of $SO(2,1)$, k is arbitrary and can vary continuously.

Diagonal Covers

There is more to say, because three-dimensional gravity is actually related to $SO(2,1)\times SO(2,1)$ gauge theory, not just to gauge theory with a single $SO(2,1)$. So we should consider covers of $SO(2,1)\times SO(2,1)$ that do not necessarily come from separate covers of the two factors.

As $SO(2,1)\times SO(2,1)$ is contractible to $SO(2)\times SO(2) = U(1)\times U(1)$, we can proceed by first analyzing the $U(1)\times U(1)$ case. We consider a $U(1)\times U(1)$ gauge theory with gauge fields A, B and a Chern-Simons action

$$I = \frac{k_L}{2\pi}\int_W A\wedge dA - \frac{k_R}{2\pi}\int_W B\wedge dB. \tag{2.9}$$

To define I in the topologically non-trivial case, we pick a four-manifold M over which everything extends and define

$$I_M = \int_M \left(\frac{k_L}{2\pi}F_A\wedge F_A - \frac{k_R}{2\pi}F_B\wedge F_B\right), \tag{2.10}$$

where F_A and F_B are the two curvatures. This is well-defined mod 2π if

$$I_X = \int_X \left(\frac{k_L}{2\pi} F_A \wedge F_A - \frac{k_R}{2\pi} F_B \wedge F_B \right) \tag{2.11}$$

is a multiple of 2π for any $U(1) \times U(1)$ gauge field over a closed four-manifold X.

In $U(1) \times U(1)$ gauge theory, the charge lattice is generated by charges $(1,0)$ and $(0,1)$, and the cohomology classes $x = F_A/2\pi$ and $y = F_B/2\pi$ are integral. For a cover of $U(1) \times U(1)$, we want to extend the charge lattice. To keep things simple, we will consider only the case that will actually be important in our application: a diagonal cover, in which one adds the charge vector $(1/n, 1/n)$ for some integer n. In this case, x and y are still integral, and their difference is divisible by n: $x = y + nz$ where n is an integral class. We have

$$I_X = 2\pi(k_L - k_R) \int_X y^2 + 2\pi k_L \int_X (n^2 z^2 + 2nyz). \tag{2.12}$$

The condition that this is a multiple of 2π for any X and any integral classes y, z is that

$$\begin{aligned} k_L &\in \begin{cases} n^{-1}\mathbb{Z} & \text{if } n \text{ is odd} \\ (2n)^{-1}\mathbb{Z} & \text{if } n \text{ is even} \end{cases} \\ k_L - k_R &\in \mathbb{Z}. \end{aligned} \tag{2.13}$$

These are also the restrictions on k_L and k_R if the gauge group is a diagonal cover of $SO(2,1) \times SO(2,1)$ with action (2.3). For example, the group $SO(2,2)$ is a double cover of $SO(2,1) \times SO(2,1)$, and this cover corresponds to the case $n = 2$ of the above discussion. So for $SO(2,2)$ gauge theory, the appropriate restriction on the Chern-Simons levels is

$$\begin{aligned} k_L &\in \frac{1}{4}\mathbb{Z} \\ k_L - k_R &\in \mathbb{Z}. \end{aligned} \tag{2.14}$$

For a general n-fold diagonal cover, one should use (2.13).

It is also possible to form a "universal diagonal cover," corresponding roughly to the limit $n \to \infty$ in the above formulas. With this gauge group, there is no restriction on k_L, but $k_L - k_R$ is an integer. In terms of three-dimensional gravity, to which we return next, this corresponds to letting ℓ/G be a freely variable parameter, while the gravitational Chern-Simons coupling k' defined in eqn. (2.2) is an integer. As explained in section 1.3, although this is the state of affairs classically in three-dimensional gravity, it cannot be the correct answer quantum mechanically.

2.2. Comparison To Three-Dimensional Gravity

So far, we have understood the appropriate gauge theory normalizations for the Chern-Simons action

$$\begin{aligned} I =& k_L I_L + k_R I_R \\ =& \frac{k_L}{4\pi}\int \mathrm{tr}\left(A_L\wedge dA_L+\frac{2}{3}A_L\wedge A_L\wedge A_L\right)-\frac{k_R}{4\pi}\int \mathrm{tr}\left(A_R\wedge dA_R+\frac{2}{3}A_R\wedge A_R\wedge A_R\right). \end{aligned} \tag{2.15}$$

Our next step will be to express A_L and A_R, which are gauge fields of $SO(2,1)\times SO(2,1)$ (or a covering group) in terms of gravitational variables, and thereby determine the constraints on the gravitational couplings. We have

$$I=\frac{k_L+k_R}{2}\,(I_L-I_R)+(k_L-k_R)\frac{(I_L+I_R)}{2}. \tag{2.16}$$

The term in (2.16) proportional to I_L-I_R will gave the Einstein-Hilbert action (1.1), while the term proportional to $(I_L+I_R)/2$ is equivalent to the gravitational Chern-Simons coupling (2.2) with coefficient $k'=k_L-k_R$.

The spin connection $\omega^{ab}=\sum_i dx^i\omega_i^{ab}$ is a one-form with values in antisymmetric 3×3 matrices. The vierbein is conventionally a one-form valued in Lorentz vectors, $e^a=\sum_i dx^i\,e_i^a$. The metric is expressed in terms of e in the usual way, $g_{ij}dx^i\otimes dx^j=\sum_{ab}\eta_{ab}e^a\otimes e^b$, where $\eta=\mathrm{diag}(-1,1,1)$ is the Lorentz metric; and the Riemannian volume form is $d^3x\sqrt{g}=\frac{1}{6}\epsilon_{abc}e^a\wedge e^b\wedge e^c$, where ϵ_{abc} is the antisymmetric tensor with, say, $\epsilon_{012}=1$. All this has an obvious analog in any dimension. However, in three dimensions, a Lorentz vector is equivalent to an antisymmetric tensor; this is the fact that makes it possible to relate gravity and gauge theory. It is convenient to introduce ${}^*e_{ab}=\epsilon_{abc}e^c$, which is a one-form valued in antisymmetric matrices, just like ω. We raise and lower local Lorentz indices with the Lorentz metric η, so $\frac{1}{2}\epsilon^{abc}\epsilon_{bcd}=-\delta^a_d$, and $e^c=-\frac{1}{2}\epsilon^{abc}\,{}^*e_{bc}$.

We can combine ω and ${}^*e$ and set $A_L=\omega-{}^*e/\ell$, $A_R=\omega+{}^*e/\ell$. A small computation gives

$$I_L-I_R=-\frac{1}{\pi\ell}\int \mathrm{tr}\,{}^*e(d\omega+\omega\wedge\omega)-\frac{1}{3\pi\ell^3}\int \mathrm{tr}\,({}^*e\wedge{}^*e\wedge{}^*e). \tag{2.17}$$

In terms of the matrix-valued curvature two-form $R^{ab}=(d\omega+\omega\wedge\omega)^{ab}=\frac{1}{2}\sum_{ij}dx^i\wedge dx_jR^{ab}_{ij}$, where R^{ab}_{ij} is the Riemann tensor, and the metric tensor g, this is equivalent to

$$I_L-I_R=\frac{1}{\pi\ell}\int d^3x\sqrt{g}\left(R+\frac{2}{\ell^2}\right). \tag{2.18}$$

Remembering the factor of $(k_L+k_R)/2$ in (2.16), we see that this agrees with the Einstein-Hilbert action (1.1) precisely if

$$k_L + k_R = \frac{\ell}{8G}. \tag{2.19}$$

The central charge of the boundary conformal field theory was originally computed by Brown and Henneaux [21] for the case that the gravitational Chern-Simons coupling $k' = k_L - k_R$ vanishes. In this case, we set $k = k_L = k_R = \ell/16G$. The formula for the central charge is $c = 3\ell/2G$, and this leads to $c = 24k$. For the case $k' = 0$, the boundary CFT is left-right symmetric, with $c_L = c_R$, so in fact $c_L = c_R = 24k$.

In general, the boundary CFT has left- and right-moving Virasoro algebras that can be interpreted (for a suitable orientation of the boundary) as boundary excitations associated with A_L and A_R respectively. So the central charges c_L and c_R are functions only of k_L and k_R, respectively. Hence the generalization of the result obtained in the last paragraph is

$$(c_L, c_R) = (24k_L, 24k_R). \tag{2.20}$$

2.3. Holomorphic Factorization

In conformal field theory in two dimensions, the ground state energy is $-c/24$. More generally, if there are separate left and right central charges c_L and c_R, the ground state energies for left- and right-movers are $(-c_L/24, -c_R/24)$. Modular invariance says that the difference between the left- and right-moving ground state energies must be an integer. In the above calculation, $(c_L - c_R)/24 = k_L - k_R$. According to (2.20), this is an integer provided that the gravitational Chern-Simons coupling $k' = k_L - k_R$ is integral.

Holomorphic factorization requires that the left- and right-moving ground state energies should be separately integral, so that there is modular invariance separately for left-moving and right-moving modes of the CFT. Thus, for holomorphic factorization, c_L and c_R must both be integer multiples of 24. This is the only constraint, since holomorphic CFT's with $c = 24$ do exist (and have been classified [23] modulo a conjecture mentioned in section 1.4).

According to (2.20), the condition for c_L and c_R to be multiples of 24 is precisely that k_L and k_R must be integers. As in our discussion of eqn. (2.7), this is the right condition if the gauge group is precisely $SO(2,1) \times SO(2,1)$ rather than a covering group.

It is possible to give an intuitive explanation of why this is the right gauge group if the boundary CFT is supposed to be holomorphically factorized. The Virasoro algebra is, of

course, infinite-dimensional, but it has a finite-dimensional subalgebra, generated by $L_{\pm 1}$ and L_0, that is a symmetry of the vacuum. It is customary to refer to the corresponding symmetry group of the vacuum as $SL(2,\mathbb{R})$, but in fact, in a holomorphic CFT, in which all energies are integers, the group that acts faithfully is really $SO(2,1)$. So a holomorphically factorized CFT has symmetry group $SO(2,1)\times SO(2,1)$, and it is natural that this is the right gauge group in a gauge theory description of (aspects of) the dual gravitational theory.

Now let us consider some other possible gauge groups. One possibility is to take a double cover of each factor of $SO(2,1)\times SO(2,1)$, taking the gauge group to be $SL(2,\mathbb{R})\times SL(2,\mathbb{R})$. The appropriate restriction on the gauge theory couplings was determined in (2.8) (where we should set $n=2$) and is that k_L and k_R take values in $\frac{1}{4}\mathbb{Z}$. Hence the central charges c_L and c_R are multiples of 6. In particular, $SL(2,\mathbb{R})\times SL(2,\mathbb{R})$ gauge theory would allow us to consider values of the couplings that contradict modular invariance of the boundary CFT.

There is perhaps a more intuitive argument suggesting that $SL(2,\mathbb{R})\times SL(2,\mathbb{R})$ is not the right group to consider. If the gauge group is $SL(2,\mathbb{R})\times SL(2,\mathbb{R})$, then upon taking the two-dimensional representation of one of the $SL(2,\mathbb{R})$'s, we get a two-dimensional real vector bundle over W which generalizes what in classical geometry is the spin bundle. Thus, this would be a theory in which, in the classical limit, W is a spin manifold, endowed with a distinguished spin structure (or even two of them). That is appropriate in a theory with fermions, but not, presumably, in a theory of pure gravity.

We similarly lose modular invariance and have difficult to interpret geometric structures if we consider other non-diagonal covers of $SO(2,1)\times SO(2,1)$. So let us discuss the diagonal covers of $SO(2,1)\times SO(2,1)$ that were considered at the end of section 2.1. We know that the universal diagonal cover is not right, since then ℓ/G could vary continuously. This leaves the possibility of an n-fold diagonal cover for some n. The best-motivated example is perhaps the two-fold cover $SO(2,2)$, which is the symmetry of Anti de Sitter spacetime (as opposed to a cover of that spacetime). In this case, according to (2.14), k_L and k_R can take values in $\mathbb{Z}/4$, as long as $k'=k_L-k_R$ is integral. For the boundary CFT, this means that c_L and c_R can be multiples of 6 (with their difference a multiple of 24). For example, let us consider a hypothetical CFT of $(c_L,c_R)=(6,6)$. The ground state energies are $(-1/4,-1/4)$. As these values are not integers, the symmetry group of the ground state is not $SO(2,1)\times SO(2,1)$, but a four-fold diagonal cover, which is a double

cover of the gauge group $SO(2,2)$ that was assumed. Such a CFT cannot be holomorphically factorized, since the left- and right-moving ground state energies are not integers; it cannot even be holomorphically factorized up to a phase.[10] The structure is considerably more complicated than in the holomorphically factorized case. Similar remarks apply to other diagonal covers.

Pragmatically, the most important argument against trying to describe three-dimensional gravity via a cover of $SO(2,1) \times SO(2,1)$ may be simply the fact that no good candidates are known. For example, no especially interesting bosonic CFT (as opposed to a superconformal field theory) seems to be known at $c = 6, 12$, or 18, which are values that we would expect if we use the gauge group $SO(2,2)$. By contrast, at $c = 24$, which is natural for $SO(2,1) \times SO(2,1)$, the Frenkel-Lepowsky-Meurman monster theory is a distinguished candidate, as we mentioned in section 1.4 and will explain in more detail in section 3.

The simplest possible hypothesis is that the right gauge group to consider in studying three-dimensional pure gravity is $SO(2,1) \times SO(2,1)$, with integral k_L and k_R and holomorphic factorization of the boundary CFT. It would be highly unnatural to overlook the fact that $SO(2,1) \times SO(2,1)$ gauge theory leads to precisely the values of the central charge at which the drastic simplification of the boundary CFT known as holomorphic factorization is conceivable. Moreover, the fact that the classical action (2.15) of the gauge theory is a sum of decoupled actions for A_L and A_R – related respectively to left- and right-moving modes of the boundary CFT – is a hint of holomorphic factorization of the boundary theory.

At any rate, regardless of whether they give the whole story, it does seem well-motivated to look for holomorphically factorized CFT's with central charges multiples of 24 that are dual to three-dimensional pure gravity at special values of the cosmological constant. That will be our main focus in the rest of this paper.

[10] The left-moving partition function would have to be $\Phi/\Delta^{1/4}$, where $\Delta = \eta(q)^{24}$ is the discriminant, η being the Dedekind eta function, and Φ is a modular form of weight 3. The power of Δ was determined to get the right ground state energy; modular invariance implies that the modular weight of Φ must equal that of $\Delta^{1/4}$. As there is no modular form of weight 3, such a theory does not exist. Even at $(c_L, c_R) = (12, 12)$, it is not possible to have holomorphic factorization up to a phase. To get modular invariance and the right ground state energy, the left-moving partition function would have to be $E_6/\Delta^{1/2}$, where E_6 is the Eisenstein series of weight 6. However, the coefficients in the q-expansion of this function are not positive.

2.4. Interpretation

A few further words of interpretation seem called for.

We do *not* claim that three-dimensional gravity is equivalent, nonperturbatively, to Chern-Simons gauge theory. Some objections to this idea were described in section 1.1. We know that Chern-Simons gauge theory is useful for perturbation theory, as was explained in that section, and we hope that it is useful for understanding some nonperturbative questions. We used the gauge theory approach to get some hints about the right values of the cosmological constant (or equivalently of the central charge) simply because it was the only tool available. We certainly do not claim to have a solid argument that the values of (c_L, c_R) suggested by $SO(2,1) \times SO(2,1)$ gauge theory are the only relevant ones.

Implicit in the gauge theory approach to quantization of the dimensionless parameter ℓ/G is an assumption that at least some of the non-geometrical states that can be described in gauge theory make sense. Indeed, when understood in classical geometry with the metric assumed to be smooth and nondegenerate, the Einstein-Hilbert action (1.1) is well-defined as a real-valued function. The topological problems that cause it to be multivalued when interpreted in gauge theory as a Chern-Simons function, and lead to quantization of ℓ/G, depend on allowing certain configurations that are natural in gauge theory, but singular in geometry because the vierbein is not invertible.

Furthermore, holomorphic factorization most likely is possible only if the path integral includes a sum over some non-geometrical configurations. In quantum gravity, we at least expect to sum over all topologies of a three-manifold W, perhaps with some fixed asymptotic behavior. The choice of topology should be expected to affect both the left- and right-movers of the boundary CFT. To achieve holomorphic factorization, there must presumably be some sort of separate topological sum for left-movers and right-movers. As this does not occur in classical geometry, it must depend on some sort of non-geometric contributions to the path integral – though these contributions may be exponentially small in many circumstances.

The Gravitational Chern-Simons Action Reconsidered

Our discussion of the quantization of the gravitational Chern-Simons coupling (2.1)

$$\Delta_0 I = \frac{k'}{4\pi} \int_W \mathrm{tr} \left(\omega \wedge d\omega + \frac{2}{3} \omega \wedge \omega \wedge \omega \right) \tag{2.21}$$

has been based entirely on gauge theory. We now want to describe the similar but slightly different answer that would come from classical differential geometry. As always, the

differences reflect the fact that the gauge theory analysis allows some non-geometrical configurations.

We first briefly restate the gauge theory analysis. ω is a connection on an $SO(2,1)$ or (in Euclidean signature) $SO(3)$ bundle over a three-manifold W. As usual, to define $\Delta_0 I$ more precisely, we pick an oriented four-manifold M of boundary W with an extension of ω over W. Then we define

$$I_M = \frac{k'}{4\pi}\int_M \mathrm{tr}\, F\wedge F. \tag{2.22}$$

If M is replaced by some other four-manifold M', and $X = M - M'$ is a four-manifold without boundary obtained by gluing together M and M', then

$$I_M - I_{M'} = \frac{k'}{4\pi}\int_X \mathrm{tr}\, F\wedge F = 2\pi k'\int_X p_1(F), \tag{2.23}$$

where $p_1(F) = (1/8\pi^2)\mathrm{tr}\, F\wedge F$ is the first Pontryagin form. In general, $\int_X p_1(F)$ can be any integer, so the condition that the indeterminacy in I_M is an integer multiple of 2π means simply that k' is an integer.

This is the right answer if the right thing to do is to simply think of ω as an $SO(3)$ or $SO(2,1)$ gauge field, ignoring its classical relation to gravity. However, in classical gravity, one can get a better answer. In classical gravity, ω is a connection on the tangent bundle TW of W. Now replace TW by $TW\oplus\epsilon$, where ϵ is a trivial real line bundle. Then ω can be regarded as a connection on $TW\oplus\epsilon$ in an obvious way, and $TW\oplus\epsilon$ extends over M as the tangent bundle of M. With this choice, (2.22) becomes

$$I_M = \frac{k'}{4\pi}\int_M \mathrm{tr}\, R\wedge R, \tag{2.24}$$

where R is the curvature form of M, and (2.23) becomes

$$I_M - I_{M'} = 2\pi k'\int_X p_1(R). \tag{2.25}$$

The effect of this is that instead of the first Pontryagin number of a general bundle over X, as in (2.23), we have here the first Pontryagin number $p_1(TX)$ of the tangent bundle of X. This number is divisible by 3, because of the signature theorem, which says that for a four-manifold X, $p_1(TX)/3$ is an integer, the signature of X. Hence, in the gravitational interpretation, the condition on k' is

$$k'\in\frac{1}{3}\mathbb{Z}. \tag{2.26}$$

There is also a variant of this. If W is a spin manifold and we are willing to define the gravitational Chern-Simons coupling (2.21) in a way that depends on the spin structure of W (this does not seem natural in ordinary gravity, but it may be natural in supergravity, which we come to in section 3.2), the condition on k' can be further relaxed. In this case, we can select M so that the chosen spin structure on W extends over M. If M' is another choice with the same property, then $X = M - M'$ is a spin manifold. But (as follows from the Atiyah-Singer index theorem for the Dirac equation) the signature of a four-dimensional spin manifold is divisible by 16. So in this situation $p_1(TX)$ is a multiple of 48, and the result for k' under these assumptions is

$$k' \in \frac{1}{48}\mathbb{Z}. \tag{2.27}$$

The Physical Hilbert Space

Now we shall discuss the meaning for gravity of the physical Hilbert space of Chern-Simons gauge theory.

In three-dimensional Chern-Simons gauge theory, one can fix a Riemann surface C and construct a Hilbert space $\mathcal{H}_C$ of physical states obtained by quantizing the given theory on C. This Hilbert space depends on the Chern-Simons couplings, so when we want to be more precise, we will call it $\mathcal{H}_C(k_L, k_R)$. In [6], it was proposed that the physical Hilbert space of three-dimensional gravity on a Riemann surface C should be obtained in essentially this way, using $SO(2,1) \times SO(2,1)$ gauge theory (or something similar, depending on whether the cosmological constant is positive, negative, or zero). Some serious objections to the claim that pure gravity and Chern-Simons gauge theory are equivalent in three dimensions were noted in section 1.1. Yet remarkable progress has made made [8,11] in understanding the quantization of $SO(2,1)$ gauge theory, and even more remarkably, in relating this quantization to Liouville theory. It therefore seems likely that the Hilbert space obtained by quantizing Chern-Simons gauge theory of $SO(2,1) \times SO(2,1)$ means something for three-dimensional gravity even if the proposal in [6] was premature. The rest of this subsection will be devoted to an attempt (not used in the rest of the paper) to reconcile the different points of view. See [39] for some useful background.

We may start by asking what we mean by the physical Hilbert space $\mathcal{H}_C$ obtained in quantum gravity by quantizing on a closed manifold C. What type of question is this space supposed to answer? Quantization on, for example, an asymptotically flat spacetime leads to a Hilbert space that can be interpreted in a relatively straightforward way, but the physical meaning of a Hilbert space obtained by quantizing on a compact spatial manifold (if such a Hilbert space can be defined at all) is not clear.

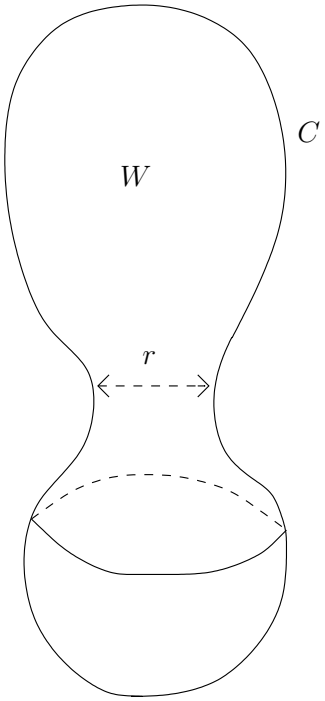


Fig. 1:

The Hartle-Hawking wavefunction Ψ is computed by integrating over three-manifolds W with a given boundary C. The dotted line labeled r denotes a path in W connecting two points in C. It is possible to vary W so that the length of this path goes to zero with no change in the geometry of C. In this limit, W becomes singular even though C is smooth. This gives a "source term" in the Wheeler-de Witt equation, because of which the Hartle-Hawking wavefunction does not obey this equation.

One line of thought that is relatively close to working is to consider the Hartle-Hawking wavefunction [40] and claim that it is a vector in $\mathcal{H}_C$. (The obvious idea that quantum mechanical probabilities would be calculated in terms of inner products of vectors in a Hilbert space $\mathcal{H}_C$ of physical states is afflicted with similar but more serious problems.) The Hartle-Hawking wavefunction is a functional of metrics on C. For every metric h on C, we define $\Psi(h)$ as the result of performing a path integral over three-manifolds W whose boundary is C and whose metric g coincides with h on the boundary. Formally, one can try to argue that $\Psi(h)$ obeys the Wheeler-de Witt equation and thus is a vector in a Hilbert space $\mathcal{H}_C$ of solutions of this equation. Moreover, one can formally match the Wheeler-de Witt equations of gravity with the conditions for a physical state in Chern-Simons gauge theory. Though many steps in these arguments work nicely, one runs into

trouble because a Riemann surface can be immersed, rather than embedded, in a three-manifold, and hence it is possible for W to degenerate without C degenerating (fig. 1). As a result, the Hartle-Hawking wavefunction does not obey the Wheeler-de Witt equation and is not a vector in $\mathcal{H}_C$.

In the case of negative cosmological constant, the boundary CFT gives a sort of cure for the problem with the Hartle-Hawking wavefunction. Instead of thinking of C as an ordinary boundary of W, we think of it as a conformal boundary at infinity. The partition function $\widehat{\Psi}(h)$ of the boundary CFT is defined by performing the path integral over all choices of W with C as conformal boundary. This is well-behaved, because, with C at conformal infinity, it is definitely embedded rather than immersed. Moreover, $\widehat{\Psi}(h)$ is a sort of limiting value of the Hartle-Hawking wavefunction. Indeed, let ϕ be a positive function on C. Then $\widehat{\Psi}(h)$ is essentially[11] the limiting value of $\Psi(e^{\phi}h)$ as $\phi\to\infty$.

This suggests that we should be able to think of $\widehat{\Psi}(h)$ as a vector in the Hilbert space $\mathcal{H}_C$ associated with three-dimensional gravity and a two-manifold C. This viewpoint can be explained more fully. The phase space of $SO(2,1)\times SO(2,1)$ Chern-Simons theory on C is the space of $SO(2,1)\times SO(2,1)$ flat connections on C. The space of $SO(2,1)$ flat connections on C has several topological components (labeled by the first Chern class, which comes in because $SO(2,1)$ is contractible to $SO(2)\cong U(1)$). One of these components, the only one that can be simply interpreted in terms of classical gravity with negative cosmological constant,[12] is isomorphic to Teichmuller space $\mathcal{T}$. Thus, this component of the classical phase space $\mathcal{M}$ is a product of two copies of $\mathcal{T}$, parametrized by a pair of points $\tau,\tau'\in\mathcal{T}$. One can quantize $\mathcal{M}$ (or at least this component of it) naively by using the standard holomorphic structure of $\mathcal{T}$. If we do this, the wavefunction of a physical state is a "function" of τ and τ' that is holomorphic in τ and antiholomorphic in τ'. (Antiholomorphy in one variable reflects the relative minus sign in the Chern-Simons action (2.15); we assume that k_L and k_R are positive.) Actually, a physical state wavefunction $\Psi(\tau,\overline{\tau}')$ is not quite a function of τ and $\overline{\tau}'$ in the usual sense, but is a form, of weights determined by k_L and k_R. So it takes values in a Hilbert space $\mathcal{H}_C(k_L,k_R)$ that depends on the Chern-Simons couplings. Such a wavefunction $\Psi(\tau,\overline{\tau}')$ is determined by its restriction

[11] One needs some renormalization in this limit; the necessary renormalization reflects the conformal anomaly. Because of this anomaly, $\widehat{\Psi}(h)$ is not quite a function only of the conformal structure of C but a function of the metric that transforms with a certain weight under conformal rescalings. As a result, the precise choice of ϕ matters, but only in a rather simple way.

[12] In the approach to quantization developed in [8,11], this is the only component considered.

to the diagonal subspace $\tau = \tau'$. Moreover, if we want to make a relation to gravity, it is natural to require that Ψ should be invariant under the diagonal action of the mapping class group on τ and τ'; this condition is compatible with restricting to $\tau = \tau'$.

Similarly, the partition function of a CFT on the Riemann surface C is a not necessarily holomorphic "function" (actually a form of appropriate weights) $\Psi(\tau, \overline{\tau})$. Being real analytic, Ψ can be analytically continued to a function $\Psi(\tau, \overline{\tau}')$ with τ' at least slightly away from τ. It does not seem to be a standard fact[13] that Ψ analytically continues to a holomorphic function on $\mathcal{T} \times \mathcal{T}$ (with invariance only under one diagonal copy of the mapping class group). However, this is true in genus 1, since the partition function can be defined as $\mathrm{Tr}\, q^{L_0} \overline{q}'^{\overline{L}_0}$, where we can take q and q' to be independent complex variables of modulus less than 1. It seems very plausible that the statement is actually true for all values of the genus, since one can move on Teichmuller space by "cutting" on a circle and inserting $q^{L_0} \overline{q}'^{\overline{L}_0}$. If so, the partition function of the CFT can always be interpreted as a vector[14] in the Chern-Simons Hilbert space $\mathcal{H}_C(k_L, k_R)$.

If we are given a theory of three-dimensional gravity, possibly coupled to other fields, the partition function of the dual CFT is a wavefunction $\Psi(\tau, \overline{\tau}')$ which, according to the conjecture just stated, is a vector in $\mathcal{H}_C(k_L, k_R)$. Any gravitational theory of the same central charges leads to another vector in the same space.

From this point of view, it seems that we should not claim, as was done in [6], that $\mathcal{H}_C(k_L, k_R)$ is a space of physical states that are physically meaningful in pure three-dimensional gravity. Rather, a particular bulk gravitational theory, such as pure gravity, gives rise to a particular dual CFT whose partition function gives a definite vector in $\mathcal{H}_C(k_L, k_R)$. Another gravitational theory, perhaps with matter fields, whose dual CFT has the same values of the central charges, will lead to a dual partition function that is another vector in the same space. Thus, $\mathcal{H}_C(k_L, k_R)$ is in a sense a universal target for gravitational theories – with arbitrary matter fields – of given central charges.

We have formulated this for a particular Riemann surface C, but in either the gravitational theory or the dual CFT, C can vary and there is a nice behavior when C degenerates. So it is more natural to think of this as a structure that is defined for all Riemann surfaces. In conformal theory, this perspective is described in [41].

[13] However, G. Segal has obtained results in this direction.

[14] It may be necessary here to extend $\mathcal{H}_C(k_L, k_R)$ to a space of forms on $\mathcal{T} \times \mathcal{T}$ that are invariant under the action of the mapping class group but not necessarily square-integrable.

2.5. Analog For Supergravity

Here we will discuss the extension of some of these ideas to three-dimensional supergravity.

We consider primarily minimal supergravity, corresponding to the case that the boundary CFT has $\mathcal{N}=1$ supersymmetry for left-movers or right-movers or perhaps both. Thus, for, say, the left-movers, the Virasoro algebra is replace by an $\mathcal{N}=1$ super-Virasoro algebra. The symmetry algebra generated by $L_{\pm 1}$ and L_0 is extended (in the Neveu-Schwarz sector) to a superalgebra that also includes the fermionic generators $\mathcal{G}_{\pm 1/2}$. This is the Lie superalgebra of the supergroup $OSp(1|2)$, whose bosonic part is $Sp(2,\mathbb{R})$, or equivalently $SL(2,\mathbb{R})$. In particular, since the operators $\mathcal{G}_{\pm 1/2}$ transform in the two-dimensional representation of $SL(2,\mathbb{R})$, the relevant group is definitely $SL(2,\mathbb{R})$ (or possibly a covering of it), not its quotient $SO(2,1)$.

For definiteness, consider a two-dimensional CFT with $(0,1)$ supersymmetry, that is, with $\mathcal{N}=1$ supersymmetry for right-movers and none for left-movers. Then left-movers have an ordinary Virasoro symmetry and right-movers have an $\mathcal{N}=1$ super-Virasoro symmetry. Such a theory can be dual to a three-dimensional supergravity theory, which classically can be described by a Chern-Simons gauge theory in which the gauge supergroup is $SO(2,1)\times OSp(1|2)$, or possibly a cover thereof. For brevity, we will here assume that the gauge group is precisely $SO(2,1)\times OSp(1|2)$. The action is the obvious analog of (2.15):

$$\begin{aligned} I =& k_L I_L + k_R I_R \\ =& \frac{k_L}{4\pi}\int \mathrm{tr}\left(A_L\wedge dA_L+\frac{2}{3}A_L\wedge A_L\wedge A_L\right)-\frac{k_R}{4\pi}\int \mathrm{str}\left(A_R\wedge dA_R+\frac{2}{3}A_R\wedge A_R\wedge A_R\right). \end{aligned} \tag{2.28}$$

Here A_L is an $SO(2,1)$ gauge field, A_R is an $OSp(1|2)$ gauge field, and str is the supertrace in the adjoint representation of $OSp(1|2)$.

We want to generalize the analysis of section 2.1 to determine the allowed values of k_L and k_R. There is actually almost nothing to do. A_L is simply an $SO(2,1)$ gauge field, so k_L must be an integer. As for A_R, we can for topological purposes replace the supergroup $OSp(1|2)$ by its bosonic reduction $SL(2,\mathbb{R})$, since the fermionic directions are infinitesimal and carry no topology. So we can borrow the result of (2.8):

$$\begin{aligned} k_L &\in \mathbb{Z} \\ k_R &\in \frac{1}{4}\mathbb{Z}. \end{aligned} \tag{2.29}$$

We still have $(c_L, c_R) = (24k_L, 24k_R)$, since the Brown-Henneaux computation of the central charge depends only on the bosonic part of the action. So c_L must be a multiple of 24, as before, but now it seems that c_R should be a multiple of 6.

This is not the result that one might hope for, because holomorphic factorization in $\mathcal{N} = 1$ superconformal field theories requires that c_R should be a multiple of 12, not 6. So half of the seemingly allowed values of c_R are difficult to interpret in the spirit of this paper. Replacing $SO(2,1) \times OSp(1|2)$ by a covering group would only make things worse, as in the bosonic case.

There are many conceivable ways to interpret this result, including the possibility that our assumptions have been too optimistic. However, one additional possibility seems worthy of mention here. Part of the structure of a superconformal field theory is that there are Ramond-sector vertex operators. They introduce a "twist" in the supercurrent, which has a monodromy of -1 around a point at which a Ramond vertex operator is inserted. Let us assume that the superconformal dual of three-dimensional supergravity should be a theory in which Ramond vertex operators make sense. What is the gravitational dual of an insertion of such a vertex operator?

As in fig. 2, points in the conformal boundary of spacetime at which a Ramond vertex operator is inserted must be connected in the bulk by lines – such as the line labeled L in the figure – around which the fermionic fields of $OSp(1|2)$ (that is, the gravitinos) have a monodromy -1. It is plausible to interpret these lines, which we will call Ramond lines, as world-lines of Ramond-sector black holes. A spacetime history containing a black hole trajectory is of course more complicated than a spacetime with a simple line drawn in it. But for our present purposes, which are purely topological, the difference may not be important. There can also be Neveu-Schwarz sector black holes (and black holes in purely bosonic gravity), but we are about to discuss an effect which seems special to Ramond-sector black holes.

So now we have a new problem. We want to define

$$I_R = \frac{1}{4\pi} \int_W \mathrm{str} \left(A_R \wedge dA_R + \frac{2}{3} A_R \wedge A_R \wedge A_R \right) \tag{2.30}$$

in the presence of a Ramond world-line L on W. In the absence of the Ramond line, we know that I_R in $OSp(1|2)$ gauge theory is defined modulo $4 \cdot 2\pi$, which is why as stated in (2.29), k_R can be a multiple of $1/4$. What happens in the presence of the Ramond line?

Rather than solving a new topological problem, we can take the following shortcut. Let W' be a new three-manifold obtained by taking a double cover of W branched over the

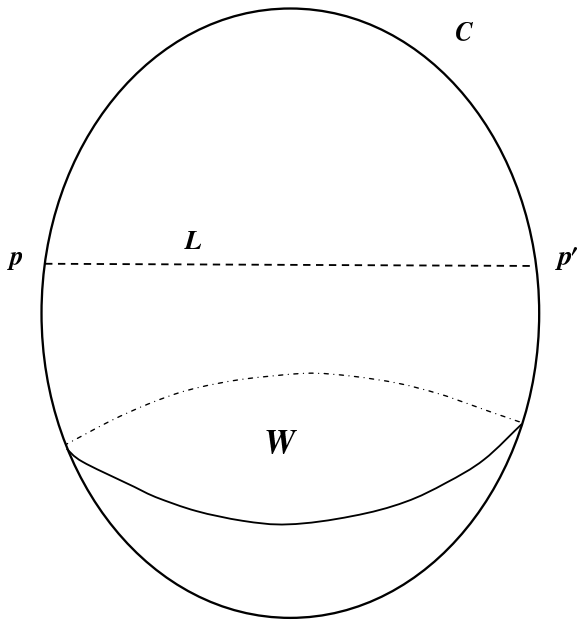


Fig. 2:

C is a Riemann surface with a pair of insertions of Ramond-sector vertex operators at points labeled p and p'. They are connected by a line L that runs in a three-manifold W whose boundary is C. Supergravity on W contributes to conformal field theory on C. The fermion fields of $OSp(1|2)$ receive a minus sign in monodromy around L.

line L. We let $I_R(W)$ be the action (2.30) and $I_R(W')$ be the corresponding action for the gauge field A_R pulled back to W'. When pulled back to W', the singularity of A_R along the Ramond line disappears, so $I_R(W')$ is defined modulo $4 \cdot 2\pi$. There is no better way to define $I_R(W)$ in the presence of a Ramond line than to say that $I_R(W) = I_R(W')/2$. So $I_R(W)$ is defined modulo $2 \cdot 2\pi$.

This means that k_R should be a multiple of $1/2$, not $1/4$. So in other words, if including Ramond lines is the right thing to do, we get

$$\begin{aligned} k_L &\in \mathbb{Z} \\ k_R &\in \frac{1}{2}\mathbb{Z}, \end{aligned} \tag{2.31}$$

and hence c_L and c_R are multiples of 24 and 12, respectively.

Unfortunately, the above "derivation" is little more than a scenario to try to justify the answer that we hoped for. However, a good pragmatic reason to focus on the case

that c_R is a multiple of 12 is that there are interesting candidate superconformal field theories (SCFT's) in that case, as we discuss in section 3. There are no obvious interesting candidates at $c_R = 6, 18$, etc.

In the supersymmetric case it is convenient to express the Chern-Simons coupling k as $k = k^*/2$, where we will focus on the case that k^* is an integer. In terms of k^*, the central charge is $c = 12k^*$.

Some Generalizations

We conclude this section by briefly mentioning some simple generalizations.

First of all, $(1,1)$ supersymmetry in two dimensions, with $\mathcal{N} = 1$ super-Virasoro symmetry for both left-movers and right-movers, is dual to three-dimensional supergravity theories related to $OSp(1|2) \times OSp(1|2)$ Chern-Simons gauge theory. If one wants the left- or right-movers to have more than $\mathcal{N} = 1$ supersymmetry, one simply replaces $OSp(1|2)$ by an appropriate supergroup with more fermionic generators. For example, $OSp(2|2)$ is related to $\mathcal{N} = 2$ supersymmetry, $PSU(2|2)$ is related to what is usually called $\mathcal{N} = 4$ supersymmetry (with the "small" $\mathcal{N} = 4$ superconformal algebra), and $OSp(r|2)$ with $r > 2$ is related to theories with "large" $\mathcal{N} = k$ superconformal algebras. (The most widely studied case of these algebras is $r = 4$; for example, see [42].)

It is also possible to consider other extended chiral algebras, apart from superconformal algebras. For example, one can start in three dimensions with an $SL(3,\mathbb{R})$ Chern-Simons gauge theory, which plausibly may be related to some sort of three-dimensional W_3 gravity theory in much the same (not fully understood) way that $SL(2,\mathbb{R})$ Chern-Simons theory is related to ordinary three-dimensional gravity. A dual CFT would very likely then have a W_3 chiral algebra. An analogous statement plausibly holds for many extended chiral algebras.

3. Partition Functions

In this section, we will determine what we propose to be the exact spectrum of physical states of three-dimensional gravity or supergravity with negative cosmological constant, in a spacetime asymptotic at infinity to Anti de Sitter space. Equivalently, we will determine the genus one partition function of the dual CFT.

In all cases, we work at the values of ℓ/G at which holomorphic factorization is possible. These values were related to gauge theory in section 2. We assume holomorphic

factorization, and mainly consider only the holomorphic sector of the theory. The full partition function is the product of the function we determine and a similar antiholomorphic function. These partition functions have been studied before [31,32], with different motivation.

3.1. The Bosonic Case

We begin with the bosonic case. What are the physical states of pure gravity in a spacetime asymptotic at infinity to AdS_3?

Since there are no gravitational waves in the theory, the only state that is obvious at first sight is the vacuum, corresponding in the classical limit to Anti de Sitter space. In a conformal field theory with central charge $c = 24k$, the ground state energy is $L_0 = -c/24 = -k$. The contribution of the ground state $|\Omega\rangle$ to the partition function $Z(q) = \mathrm{Tr}\, q^{L_0}$ is therefore q^{-k}.

Of course, there is more to the theory than just the ground state. According to Brown and Henneaux [21], a proper treatment of the behavior at infinity leads to the construction of a Virasoro algebra that acts on the physical Hilbert space. The Virasoro generators L_n, $n \geq -1$ annihilate $|\Omega\rangle$, but by acting with $L_{-2}, L_{-3}, \ldots$, we can make new states of the general form $\prod_{n=2}^{\infty} L_{-n}^{s_n}|\Omega\rangle$, with energy $-k+\sum_n n s_n$. (We assume that all but finitely many of the s_n vanish.) If these are the only states to consider, then the partition function would be

$$Z_0(q) = q^{-k} \prod_{n=2}^{\infty} \frac{1}{1-q^n}. \tag{3.1}$$

This cannot be the complete answer, because the function $Z_0(q)$ is not modular-invariant. There must be additional states such that $Z_0(q)$ is completed to a modular-invariant function.

Additional states are expected, because the theory also has BTZ black holes. The main reason for writing the present paper, after all, is to understand the role of the BTZ black holes in the quantum theory. We will assume that black holes account for the difference between the naive partition function $Z_0(q)$ and the exact one $Z(q)$. To use this assumption to determine $Z(q)$, we need to know something about the black holes.

The classical BTZ black hole is characterized by its mass M and angular momentum J. In terms of the Virasoro generators,

$$\begin{aligned} M =& \frac{1}{\ell}(L_0 + \overline{L}_0) \\ J =& (L_0 - \overline{L}_0), \end{aligned} \tag{3.2}$$

so $L_0 = (\ell M + J)/2$, $\overline{L}_0 = (\ell M - J)/2$. The classical BTZ black hole obeys $M\ell \geq |J|$, or $L_0, \overline{L}_0 \geq 0$. The BTZ black hole is usually studied in the absence of the gravitational Chern-Simons coupling, that is for $k_L = k_R = k$. Its entropy is $S = \pi(\ell/2G)^{1/2}\left(\sqrt{M\ell - J} + \sqrt{M\ell + J}\right)$. (This entropy was first expressed in terms of two-dimensional conformal field theory in [13].) With $\ell/G = 16k$ as in (2.19), this is equivalent to $S = 4\pi\sqrt{k}\left(\sqrt{L_0} + \sqrt{\overline{L}_0}\right)$. For the holomorphic sector, the entropy is therefore

$$S_L = 4\pi\sqrt{k_L L_0}, \tag{3.3}$$

and similarly for the antiholomorphic sector.

There is no classical BTZ black hole with $L_0 < 0$, and the entropy of such a black hole is zero if $L_0 = 0$. We will take this as a suggestion that quantum states corresponding to black holes exist only if $L_0 > 0$, that is $L_0 \geq 1$. This means that the exact partition function $Z(q)$ should differ from the function $Z_0(q)$ in (3.1) by terms of order q:

$$Z(q) = q^{-k}\prod_{n=2}^{\infty}\frac{1}{1-q^n} + \mathcal{O}(q). \tag{3.4}$$

Admittedly, we are here trying to squeeze more information from the classical result than is justified. But it turns out that a modular-invariant partition function of this form exists and is unique. This result (which is due to Höhn [31]) follows from the fact that the moduli space $\mathcal{M}_1$ of Riemann surfaces of genus 1 is itself a Riemann surface of genus 0, in fact parametrized by the j-function. If E_4 and E_6 are the usual Eisenstein series of weights 4 and 6, then $j = 1728E_4^3/(E_4^3 - E_6^2)$. Its expansion in powers of q is

$$j(q) = q^{-1} + 744 + 196884q + 21493760q^2 + 864299970q^3 + 20245856256q^4 + \dots. \tag{3.5}$$

Actually, it is more convenient to use the function

$$J(q) = j(q) - 744 = q^{-1} + 196884q + 21493760q^2 + \dots, \tag{3.6}$$

which likewise parametrizes the moduli space.

The J-function has a pole at $q = 0$ and no other poles. The statement that J parametrizes the moduli space means precisely that any modular-invariant function can be written as a function of J. The partition function $Z(q)$ has a pole at $q = 0$, that is at $J = \infty$. This pole is of order k, and the partition function has no other poles. Any holomorphic function of J that has no singularity except at $J = \infty$ is a polynomial in J.

In the present case, as the pole in $Z(q)$ at $q = 0$ is of order k, Z must be a polynomial in J of degree k. Thus

$$Z(q) = \sum_{r=0}^{k} f_r J^r, \tag{3.7}$$

with some coefficients f_r. These $k+1$ coefficients can be adjusted in a unique fashion to ensure that $Z(q)$ takes the form (3.4), or in other words to ensure that the terms in $Z(q)$ of order q^{-n}, $n = 0, \dots, k$, coincide with the naive function $Z_0(q)$.

When we do this, we get a function that we will call $Z_k(q)$, $k = 1, 2, 3, \dots$. This function is our candidate for the generating function that counts the quantum states of three-dimensional gravity in a spacetime asymptotic to AdS_3. For example, for $k = 1$ we have simply $Z_1(q) = J(q)$, and the next few examples are

$$\begin{aligned}
Z_2(q) =& J(q)^2 - 393767 \\
=& q^{-2} + 1 + 42987520q + 40491909396q^2 + \dots \\
Z_3(q) =& J(q)^3 - 590651J(q) - 64481279 \\
=& q^{-3} + q^{-1} + 1 + 2593096794q + 12756091394048q^2 + \dots \\
Z_4(q) =& J(q)^4 - 787535J(q)^2 - 8597555039J(q) - 644481279 \\
=& q^{-4} + q^{-2} + q^{-1} + 2 + 81026609428q + 16046712924524522 76q^2 + \dots .
\end{aligned} \tag{3.8}$$

Following [31], we refer to a holomorphic CFT with $c = 24k$ and partition function $Z_k(q)$ as an extremal CFT. According to our proposals, the dual of three-dimensional gravity should be an extremal CFT.

As was already noted in the introduction, Frenkel, Lepowsky, and Meurman constructed [24,25] an extremal CFT with $k = 1$, that is, a holomorphic CFT with $c = 24$ and partition function $J(q) = Z_1(q)$. They also conjectured its uniqueness. If that conjecture as well as the ideas in the present paper are correct, then the FLM theory must be the dual to quantum gravity for $k = 1$. Unfortunately, as also noted in the introduction, for $k > 1$, extremal CFT's are not known, though their possible existence has been discussed in the literature [31-33] for reasons not related to three-dimensional gravity. Our reasoning in this paper suggests that such theories should exist and be unique for each k.

The main point of the FLM construction was that their theory has as a group of symmetries the Fischer-Griess monster group $\mathbb{M}$, the largest of the sporadic finite groups. Arguably, the FLM theory is the most natural known structure with $\mathbb{M}$ symmetry. The coefficients in the q-expansion of the J-function are integers for number-theoretic reasons,

but the FLM construction gave a new perspective on why they are positive; this is a property of the partition function of any CFT. It also gave a new perspective on why these coefficients are so large; this follows from $\mathbb{M}$ symmetry, since $\mathbb{M}$ does not have small representations.

Indeed, the original clue to the FLM construction was the observation by J. McKay that the first non-trivial coefficient 196884 of the function $J(q)$ is nearly equal to the smallest dimension 196883 of a non-trivial representation of $\mathbb{M}$. (This observation was later greatly generalized [43,44].) The FLM interpretation is that 196884 is the number of operators of dimension 2 in their theory. One of these operators is the stress tensor, while the other 196883 are primary fields transforming in the smallest non-trivial representation of $\mathbb{M}$.

In our interpretation, the 196883 primaries are operators that (when combined with suitable anti-holomorphic factors) create black holes. It is illuminating to compare the number 196883 to the Bekenstein-Hawking formula. An exact quantum degeneracy of 196883 corresponds to an entropy of $\ln 196883 \cong 12.19$. By contrast, the Bekenstein-Hawking entropy at $k = 1$ and $L_0 = 1$ is $4\pi \cong 12.57$. We should not expect perfect agreement, because the Bekenstein-Hawking formula is derived in a semiclassical approximation which is valid for large k.

Agreement improves rapidly if one increases k. For example, at $k = 4$, and again taking $L_0 = 1$, the exact quantum degeneracy of primary states is 81026609426, according to eqn. (3.8). (Two of the states at this level are descendants.) This corresponds to an entropy $\ln 81026609426 \cong 25.12$, compared to the Bekenstein-Hawking entropy $8\pi \cong 25.13$. Shortly we will compute the entropy in the large k limit.

Our interpretation is that a primary state $|\Lambda\rangle$ represents a black hole, while a descendant $\prod_{n=1}^{\infty} L_{-n}^{s_n} |\Lambda\rangle$ describes a black hole embellished by boundary excitations. These boundary excitations are the closest we can come in $2+1$ dimensions to the gravitational waves that a black hole can interact with in a larger number of dimensions. If this interpretation is correct, then for an exact count of black hole states, we should count primaries only. However, as the above examples illustrate, in practice this issue has only a very slight effect on the black hole degeneracies. The boundary excitations contribute to the entropy an amount that is independent of k and thus negligible in the regime where we compare to the Bekenstein-Hawking entropy, and also, numerically, negligible even for small k. The separation between the black hole and the boundary excitations is best-motivated for large k.

Alternative Formula

Now we will present an alternative formula for the partition functions $Z_k(q)$. This formula avoids the large coefficients involved in writing the Z_k as a polynomial in J, and will make some properties of the Z_k more manifest.

Let $q = \exp(2\pi i\tau)$ with τ in the upper half plane. If p is a prime number, then the Hecke operators acting on functions $F(\tau)$ are defined essentially by

$$T'_p F(\tau) = F(p\tau) + \sum_{b=0}^{p-1} F((\tau + b)/p). \tag{3.9}$$

(What we call T'_p is p times the Hecke operator T_p as usually defined. See [45] for an introduction to Hecke operators and [29,30] for their use in the present subject.) More generally, for any positive integer t, we define

$$T'_t F(\tau) = \sum_{d|t} \sum_{b=0}^{d-1} F((t\tau + bd)/d^2). \tag{3.10}$$

For $t = 0$, we define $T'_0 F(\tau) = 1$. If $F(\tau)$ is modular-invariant, then so is $T'_t F(\tau)$. The definition of Hecke operators generalizes naturally to modular forms, but we will not need this.

An immediate consequence of the definition is that, for any $F(\tau)$ that is invariant under $\tau \to \tau + 1$, and so has a Laurent expansion in powers of $q = \exp(2\pi i\tau)$, the coefficients in the q-expansion of $T'_n F(\tau)$ are linear combinations with positive integer coefficients of the q-expansion coefficients of $F(\tau)$. The FLM construction shows that the q-expansion coefficients of the J-function are non-negative integers and the coefficients of positive powers of q are dimensions of non-trivial (reducible) monster representations. Hence the same is true of $T'_n J$.

An important special case of how the Hecke operators act on q-expansion coefficients is that if $F(\tau) = q^{-1} + \mathcal{O}(q)$, then

$$T'_n F(\tau) = q^{-n} + \mathcal{O}(q). \tag{3.11}$$

Now we can give an alternative description of the partition functions $Z_k(\tau)$. We recall that these functions are uniquely determined by being modular-invariant and agreeing up to order q with the function $Z_0 = q^{-k} \prod_{n=2}^{\infty} (1 - q^n)^{-1}$. To find a function with these properties, we simply expand

$$Z_0 = \sum_{r=-k}^{\infty} a_r q^r, \tag{3.12}$$

and then we let

$$Z_k(\tau) = \sum_{r=0}^{k} a_{-r} T'_r J(\tau). \tag{3.13}$$

For example,

$$Z_2(\tau) = (T'_2 + T'_0)J(\tau) = J(2\tau) + J(\tau/2) + J((\tau+1)/2) + 1. \tag{3.14}$$

This representation makes it clear that, just like the $T'_r J$, the functions Z_k share some properties with the J-function: the q-expansion coefficients are non-negative, and the coefficients of positive powers of q are dimensions of non-trivial monster representations.

This makes it tempting to speculate that the CFT's dual to three-dimensional gravity have $\mathbb{M}$ symmetry for all k. If so, the $\mathbb{M}$ symmetry is invisible in classical General Relativity and acts on the microstates of black holes. (However, the analogous conjecture for supergravity, which would involve the Conway group, appears to be untrue, as we will see in section 3.3.)

As an illustration of the usefulness of the expression for the partition function in terms of Hecke operators, we will use it to compare with the Bekenstein-Hawking entropy. We consider the semiclassical limit $k, L_0 \to \infty$, with $r = L_0/k$ positive and fixed. We take $r = p/q$ to be a rational number and assume that k is divisible by q, so that L_0 is an integer n. If we write

$$J(q) = \sum_{m=-1}^{\infty} c_m q^m, \tag{3.15}$$

then a formula of Petersson and Rademacher gives the asymptotic behavior

$$\ln c_m \sim 4\pi\sqrt{m} - \frac{3}{4}\ln m - \frac{1}{2}\ln 2 + \dots. \tag{3.16}$$

Let

$$Z_k(\tau) = \sum_{n=-k}^{\infty} b_{k,n} q^n. \tag{3.17}$$

We want to determine the behavior of $b_{k,n}$ for large k and n with $r = n/k$ fixed. To evaluate the partition function using the formula (3.13), let us first look at the contribution from $r = k$, that is $T'_k J(\tau)$. In evaluating $T'_k J$ from the definition (3.10), the dominant term (for large k and n) is the term with $d = k$. This contribution to $b_{k,n}$ is $b^0_{k,n} \sim k c_{kn}$, and hence, using the Petersson-Rademacher formula,

$$\ln b^0_{k,n} \sim 4\pi\sqrt{kn} + \frac{1}{4}\ln k - \frac{3}{4}\ln n - \frac{1}{2}\ln 2 + \dots. \tag{3.18}$$

The first term is the Bekenstein-Hawking entropy (3.3). The additional terms in (3.18), as well as the remaining contributions in (3.13) with $r < k$ that we have omitted in deriving (3.18), do not modify the Bekenstein-Hawking formula in the limit of large k, fixed n/k, but give interesting subleading corrections that will be studied elsewhere [46].

Another method to get similar results (expressing black hole degeneracies in terms of coefficients of the singular part of the partition function) is the Farey tail expansion [34]. See also [47].

The FLM Construction

Consider a holomorphic CFT with $c = 24$. If the number of primary fields of dimension 1 is s, then the expansion of the partition function near $q = 0$ begins $Z(q) = q^{-1} + s + \mathcal{O}(q)$, and modular invariance implies that the genus 1 partition function is precisely $Z(q) = J(q) + s = q^{-1} + s + 196884q + \ldots$. Of the 196884 fields of dimension 2, one is a descendant of the identity and s are descendants of dimension 1 primaries. Hence there are $196883 - s$ primary fields of dimension 2, too few to furnish a non-trivial representation of the monster group unless $s = 0$.

For this reason, Frenkel, Lepowsky, and Meurman [24] aimed to construct a holomorphic CFT of $c = 24$ with no primary fields of dimension 1, hoping that it would have monster symmetry. Their construction was made as follows. The Leech lattice is an even, unimodular lattice of rank 24 with no vector of length squared less than 4. One can construct a holomorphic theory of $c = 24$ by compactifying 24 chiral bosons X_i, $i = 1, \ldots, 24$ using any even unimodular lattice. If one uses the Leech lattice, then because it contains no vector of length squared less then 4, a primary field of the form $\exp(ip \cdot X)$ has dimension at least 2, as desired. However, this lattice theory has 24 primary fields ∂X_i of dimension 1. Aiming to eliminate those fields, FLM considered an orbifold[15] by the $\mathbb{Z}_2$ symmetry $X_i \to -X_i$ which eliminates the dimension 1 primaries. As it turns out, this particular orbifold preserves modular invariance, does not add dimension 1 primaries in twisted sectors, and gives the hoped-for theory with monster symmetry.

FLM conjectured that the model they constructed is the unique holomorphic CFT with partition function J. If this conjecture is correct, it implies that many other constructions give the same theory. There are [49] many orbifolds of the Leech lattice theory that give

[15] Their work preceded the general study of orbifolds in string theory [48], though some related stringy constructions, such as the Ramond and Neveu-Schwarz sectors of string theory, were already known.

holomorphic CFT's with no primary of dimension 1. According to the FLM uniqueness conjecture, all of these theories are isomorphic to their monster theory. Moreover, the Leech lattice theory itself can be obtained as an orbifold theory starting from another even unimodular lattice of rank 24, so other lattices can be used as starting points as well.

Aiming to imitate the FLM construction for $k = 2$, one might start with an even unimodular rank 48 lattice with no vector of length squared less than 6. Three such lattices are known; two are described in [50] and a third has been constructed more recently [51]. (There is no reason to believe that these are the only three rank 48 lattices with this property; there may be a vast number of them.) In the theory of 48 free chiral bosons X_i compactified using such a lattice, a primary field $\exp(ip \cdot X)$ has dimension at least 3, but there are dimension 1 primaries ∂X_i and dimension 2 primaries $\partial X_i \partial X_j - \frac{1}{48}\delta_{ij}\partial X \cdot \partial X$. A simple $\mathbb{Z}_2$ orbifold eliminates the dimension 1 primaries, as in the FLM case, but not the dimension 2 primaries. One may attempt to find a more complicated orbifolding construction to remove the unwanted primaries. Though all three lattices have interesting discrete symmetry groups, it appears that there is no anomaly-free subgroup that removes all dimension 2 primaries.[16] Many other orbifolds can be considered, such as the symmetric product of k copies of the $k = 1$ monster theory, but it appears difficult to remove all primaries of low dimension.

Optimistically speaking, this situation might be compared to current algebra of a simply-laced compact Lie group G. At level 1, one has the Frenkel-Kac-Segal construction of current algebra of G via free bosons. At higher integer level, the theory still exists, but generically has no equally straightforward realization in terms of free fields. Perhaps the situation is somewhat similar for extremal CFT's.

Further Remarks

We have emphasized here the partition function, because it is what we can determine for general k. However, if one can actually describe the relevant CFT – as conjecturally we can at $k = 1$ via the FLM construction – this gives much more than a partition function. In this case, by computing matrix elements of primary fields, we get a detailed description of the black hole quantum mechanics.

Above 2+1 dimensions, a black hole can form out of radiation and ordinary matter. In pure gravity in $2+1$ dimensions, the only thing that a black hole can form from is, roughly

[16] I am grateful for assistance from G. Nebe in investigating this question.

speaking, smaller black holes. To be more precise, let Φ_i, $i = 1, \ldots, w$ be primary fields of dimension not much greater than k that individually could, in acting on the vacuum, create black holes of fairly small mass. A product of such fields acting on the vacuum at prescribed points

$$\prod_{i=1}^{w} \Phi_i(z_i)|\Psi\rangle \tag{3.19}$$

may create a black hole of large mass, that is, a primary state of large energy or a descendant of such a state. Evaluating such matrix elements is the closest analog we can find, in the present model, to describing the formation of a large mass black hole from matter. Unfortunately, at the moment, we can perform such computations only for $k = 1$, since for other values of k we do not know the CFT. Ideally, one would like to describe the CFT for all k and investigate the semiclassical limit of large k.

If our hypotheses are correct, the following may be the most significant difference between three-dimensional pure gravity and a more realistic theory of black holes in $3 + 1$ dimensions. In the present model, as the holomorphic fields are all conserved currents, and the energy levels are integers, the dynamics is integrable and periodic on a short time scale. This is certainly not expected for black holes in general. An embedding of three-dimensional gravity in a larger system, such as a string theory, would be expected to give a non-integrable deformation of the model.

A Conundrum

We have here considered the holomorphic sector of a CFT, but the CFT dual of three-dimensional gravity has both holomorphic and antiholomorphic degrees of freedom. Let us therefore discuss a puzzle (stressed by J. Maldacena) that arises when one combines the holomorphic and antiholomorphic degrees of freedom. We let $|\Omega\rangle$, $|\widetilde{\Omega}\rangle$ be the ground states in the holomorphic and antiholomorphic sectors, and let $|\Phi\rangle$ and $|\widetilde{\Phi}\rangle$ denote primary states other than the ground state. The state $|\Omega\rangle \otimes |\widetilde{\Omega}\rangle$ and its descendants correspond, according to our picture, to Anti de Sitter space and its boundary excitations. A state $|\Phi\rangle \otimes |\widetilde{\Phi}\rangle$, or a descendant thereof, corresponds to a BTZ black hole, perhaps with boundary excitations. But what do we make of states $|\Omega\rangle \otimes |\widetilde{\Phi}\rangle$ or $|\Phi\rangle \otimes |\widetilde{\Omega}\rangle$, and their descendants? Such states are trying to be Anti de Sitter space for holomorphic variables and black holes for antiholomorphic variables, or vice-versa.

In classical three-dimensional gravity, there is no satisfactory solution that has this interpretation. We can, however, see what is involved in trying to make one. First of all,

three-dimensional Anti de Sitter space AdS_3 is simply the universal cover of the $SL(2,\mathbb{R})$ group manifold, with the symmetry $SL(2,\mathbb{R})\times SL(2,\mathbb{R})$ (or rather a cover thereof) coming from the left and right action of $SL(2,\mathbb{R})$ on itself. The conformal boundary of the universal cover of AdS_3 is a cylinder $\mathbb{R}\times S^1$. We can parametrize it by $-\infty<\tau<\infty$, $0\leq\sigma\leq 2\pi$, with conformal structure described by the metric $ds^2=d\tau^2-d\sigma^2$. The BTZ black hole is the quotient of (a cover of) $SL(2,\mathbb{R})$ by a subgroup $\mathbb{Z}\subset SL(2,\mathbb{R})\times SL(2,\mathbb{R})$ that is not simply a subgroup of one factor or the other. This quotient has conformal boundary that is again a cylinder, conformally equivalent to the boundary of Anti de Sitter space itself. That is why the BTZ black hole can be regarded as an excitation of Anti de Sitter space.

To describe a state $|\Omega\rangle\otimes|\Phi\rangle$ (or its parity conjugate) at the classical level, we want a solution of Einstein's equations that is invariant under precisely one of the two factors of the $SL(2,\mathbb{R})\times SL(2,\mathbb{R})$ symmetry. This will reflect the fact that $|\Omega\rangle$ is $SL(2,\mathbb{R})$-invariant and $|\Phi\rangle$ is not. A space with just one of the $SL(2,\mathbb{R})$ symmetries is $\mathrm{AdS}_3/\mathbb{Z}$, where $\mathbb{Z}$ is a subgroup of one $SL(2,\mathbb{R})$ factor. The quotient $\mathrm{AdS}_3/\mathbb{Z}$ is a perfectly good solution of Einstein's equations in bulk, but its asymptotic behavior at infinity does not quite agree with that of AdS_3. The boundary at infinity is a cylinder, but the compact direction in the cylinder is lightlike rather than spacelike. For our approach to three-dimensional gravity in the present paper to be correct, something like this quotient $\mathrm{AdS}_3/\mathbb{Z}$ has to make sense at the quantum level, perhaps because of a small quantum correction that makes the compact direction in the cylinder effectively spacelike. Note that the metric $ds^2_\epsilon=d\tau^2-\epsilon d\sigma^2$ induces on the cylinder a conformal structure that is independent of ϵ up to isomorphism as long as $\epsilon>0$.

What we have described is an analog for black holes of a statement made in section 2.4: to ensure holomorphic factorization, the sum over topologies probably must be extended to include configurations that are difficult to interpret classically. In the present case, the analog of the sum over topologies is the sum over states with or without the presence of a black hole.

3.2. The Supersymmetric Case

Now we consider the analog for supergravity. The analysis is a little more complicated and in some ways the results are less satisfactory.

A rather similar problem has been treated by Höhn [31]. He considered not a superconformal field theory, but a more general holomorphic theory with bosonic operators of integral dimension and fermionic operators of half-integral dimension. In that case, the

partition function can have a pole at the Ramond cusp, and is given by a Laurent series in $j_\theta = K - 24$ rather than a polynomial.

We continue to assume holomorphic factorization, and we assume that the holomorphic part of the boundary CFT is an $\mathcal{N} = 1$ superconformal field theory (SCFT). This means (in whatever not well understood sense gravity is related to gauge theory) that the gauge group of the bulk theory has a factor $OSp(1|2)$. From a classical point of view, the $OSp(1|2)$ bundle over a three-manifold W endows W with a spin structure. In the AdS/CFT correspondence, W has for conformal boundary a Riemann surface C, and the spin structure on W determines one on C. One is instructed in the AdS/CFT correspondence to specify C, including its spin structure, and sum over all choices of W. (Depending on the theory, the antiholomorphic degrees of freedom may themselves be supersymmetric and endowed with a choice of spin structure.) We will assume that it is physically appropriate to specify the spin structure on C rather than summing over it (just as one specifies the complex structure of C), though at the end of section 3.3, we discuss what happens if one sums over spin structures.

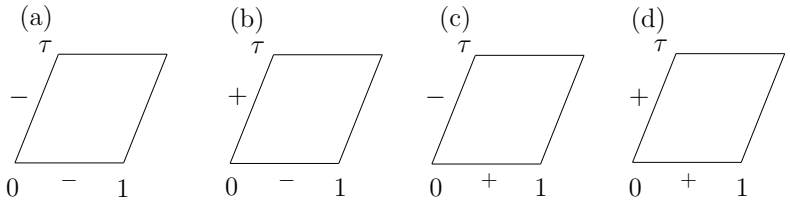


Fig. 3:

We represent a genus 1 Riemann surface C as the quotient of the complex plane by the lattice generated by complex numbers 1 and τ, where τ lies in the upper half plane $\mathcal{H}$. There are four possible spin structures, classified by whether the fermions are periodic or antiperiodic around the "horizontal" and "vertical" cycles. This is indicated by labeling the horizontal and vertical directions with a $+$ sign for periodicity or a $-$ sign for antiperiodicity.

We focus on the case that C has genus 1, in which case there are four possible spin structures (fig. 3). The path integrals for these four spin structures can be interpreted in terms of traces in two Hilbert spaces, known as the Neveu-Schwarz (NS) and Ramond (R) Hilbert spaces $\mathcal{H}_{\rm NS}$ and $\mathcal{H}_{\rm R}$, respectively.

Traces in $\mathcal{H}_{\rm NS}$ are constructed using the spin structures, shown in (a) and (b) in the figure, in which fermions are antiperiodic in the horizontal direction. If additionally the fermions are antiperiodic in the vertical direction, as in (a), then the path integral computes

$$F(\tau) = {\rm Tr}_{\mathcal{H}_{\rm NS}}\, q^{L_0}, \tag{3.20}$$

where $q = \exp(2\pi i\tau)$. We call this the NS partition function. If they are periodic in the vertical direction, as in (b), the path integral computes instead

$$G(\tau) = {\rm Tr}_{\mathcal{H}_{\rm NS}}(-1)^F q^{L_0}, \tag{3.21}$$

where $(-1)^F$ is the operator that is $+1$ on bosonic states, including the ground state of the NS sector, and -1 on fermionic states.

Traces in $\mathcal{H}_{\rm R}$ are computed using spin structures, shown in (c) and (d) in the figure, in which fermions are periodic in the horizontal direction. If the fermions are antiperiodic in the vertical direction, the path integral computes an ordinary trace

$$H(\tau) = {\rm Tr}_{\mathcal{H}_{\rm R}}\, q^{L_0}. \tag{3.22}$$

If fermions are periodic in both directions, the path integral gives

$$\chi = {\rm Tr}_{\mathcal{H}_{\rm R}}\,(-1)^F q^{L_0} = {\rm Tr}\,(-1)^F. \tag{3.23}$$

Both $H(\tau)$ and χ are severely constrained by the fact that in the R sector, one of the superconformal generators, which we will $\mathcal{G}_0$, commutes with L_0 and obeys $\mathcal{G}_0^2 = L_0$. This implies first of all that $L_0 \geq 0$ in the R sector, so $H(\tau)$ has no pole at $q = 0$. Further, the action of $\mathcal{G}_0$ implies that states of $L_0 > 0$ are paired between bosons and fermions. They cancel out of χ and make equal contributions to $H(\tau)$. It follows that χ is simply an integer, the trace of the operator $(-1)^F$ in the subspace with $L_0 = 0$ (or alternatively the index of the operator $\mathcal{G}_0$ mapping from bosonic states to fermionic ones). Moreover, if we expand $H(\tau)$ in powers of q

$$H(\tau) = \sum_{n=0}^{\infty} h_n q^n, \tag{3.24}$$

then all coefficients h_n are even except possibly h_0. In this expansion, only integral powers of q appear, since all fields obey integral boundary conditions in the R sector.

In the NS sector, the ground state energy is $-c/24$. In the holomorphically factorized case that we focus on, $c = 12k^*$ for an integer k^*, so the ground state energy is $-k^*/2$.

Since fermions in the NS sector are antiperiodic in the spatial direction, NS excitations may have either integer or half-integer energy above the ground state. The partition function therefore takes the general form

$$F(\tau) = q^{-k^*/2}(1 + aq^{1/2} + bq + cq^{3/2} + \ldots) = \sum_{m\in\mathbb{Z}/2,\, m\geq -k^*/2} f_m q^m. \tag{3.25}$$

In this expansion, the states in which $m + k/2$ is an integer are bosons, while the states in which $m + k/2$ is half-integral are fermions. That is so because, in the NS sector, bosonic fields are periodic and have integral excitation energies, while fermionic fields are half-integral and have half-integral excitation energies. $G(\tau)$ can therefore very explicitly be expressed in terms of the same coefficients:

$$G(\tau) = \sum_{m\in\mathbb{Z}/2,\, m\geq -k^*/2} (-1)^{2m+k^*} f_m q^m. \tag{3.26}$$

We have simply included a factor that is $+1$ on bosons and -1 on fermions. A more succinct way to say the same thing is that

$$G(\tau) = (-1)^{k^*} F(\tau + 1). \tag{3.27}$$

Thus, G is simply determined in terms of F. The same is true for H. Consider the modular transformation $\tau \to -1/\tau$, which has the effect of exchanging the horizontal and vertical directions in fig. 3 (with a reversal of orientation for one). This exchanges spin structures (b) and (c), as a result of which

$$H(\tau) = G(-1/\tau). \tag{3.28}$$

We can combine (3.27) and (3.28) to get

$$H(\tau) = (-1)^{k^*} F(-1/\tau + 1). \tag{3.29}$$

An important special case of this is

$$\lim_{\tau\to i\infty} H(\tau) = (-1)^{k^*} F(1). \tag{3.30}$$

The limit on the left hand side exists, because H has no pole at $q = 0$. The limit is just the coefficient h_0 in (3.24):

$$h_0 = (-1)^{k^*} F(1). \tag{3.31}$$

h_0 is the number of Ramond states of zero energy.

According to the above formulas, everything (except an integer χ, the supersymmetric index) can be expressed in terms of F. So it is useful to understand the modular properties of F. F is not invariant under $\tau \to \tau + 1$, as we have already seen. But, since all energies take values in $\mathbb{Z}/2$, F is invariant under $\tau \to \tau + 2$. In addition, since the modular transformation $\tau \to -1/\tau$ maps spin structure (a) itself, F is invariant under $\tau \to -1/\tau$. The two transformations $\tau \to \tau + 2$ and $\tau \to -1/\tau$ generate a subgroup of $SL(2,\mathbb{Z})$ that consists of 2×2 integral unimodular matrices

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \tag{3.32}$$

that are congruent mod 2 to one of

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}. \tag{3.33}$$

These matrices of course act on τ in the usual fashion, $\tau \to (a\tau + b)/(c\tau + d)$. This group is sometimes called Γ_θ; it is conjugate to the group $\Gamma_0(2)$ characterized by requiring b to be even. (A third conjugate group is characterized by the condition that c should be even. Each of these three groups can be defined as the subgroup of $SL(2,\mathbb{Z})$ that leaves fixed one of the even spin structures, that is, one of the first three shown in fig. 3.)

The quotient of the upper half plane $\mathcal{H}$ by $SL(2,\mathbb{Z})$ is a curve of genus zero with one point missing. This can be seen by studying the fundamental domain (fig. 4(a)). The missing point is called the "cusp" at $\tau = i\infty$. It is because the quotient has genus zero that Riemann surfaces of genus 1 can be parametrized by a single holomorphic function, namely $J(\tau)$. As we have seen in section 3.1, this is useful because it means that the constraints of modular invariance are equivalent to the statement that the partition function is a function of J. Similarly, the quotient of the upper half plane $\mathcal{H}$ by Γ_θ is again a curve of genus zero, this time with two missing points or cusps, as explained in fig. 4(b). This means that there is a function $K(\tau)$ such that any Γ_θ-invariant function F can be expressed as a function of K.

The two cusps of $\mathcal{H}/\Gamma_\theta$ are at $\tau \to i\infty$ and $\tau = 1$, and correspond respectively to the NS sector and the R sector. For $\tau \to i\infty$, the NS partition function $F(\tau)$ has an expansion in terms of L_0 eigenvalues in the NS sector. Likewise, the Ramond partition function $H(\tau)$ can be expanded for $\tau \to i\infty$ in L_0 eigenvalues in the Ramond sector; but according to (3.29), the behavior of $H(\tau)$ for τ near $i\infty$ is the same as the behavior of $F(\tau)$ for τ near 1.

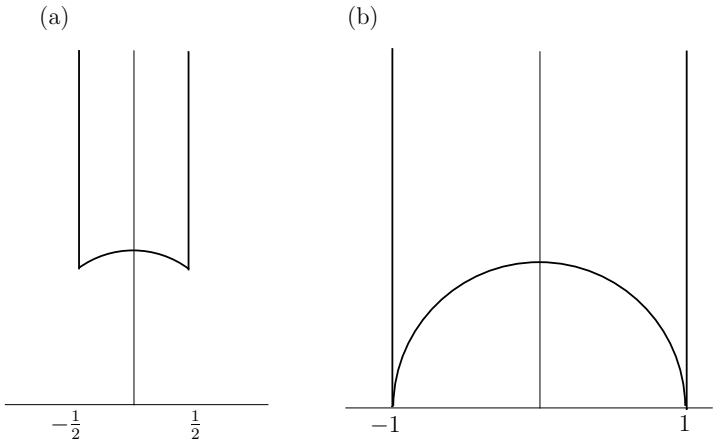


Fig. 4:

(a) A fundamental domain for the action of $SL(2,\mathbb{Z})$ on the upper half plane. By the action of $\tau \to \tau + 1$, one can take $|\mathrm{Re}\,\tau| \leq 1/2$, and by the action of $\tau \to -1/\tau$, one can take $|\tau| \geq 1$. The $SL(2,\mathbb{Z})$ action identifies the left and right hand halves of the boundary of the fundamental domain, making what topologically is a Riemann surface of genus zero with one missing point at $\tau = i\infty$. This point is called the cusp. (b) The analogous fundamental domain for the action of Γ_θ. Here the symmetry $\tau \to \tau + 2$ lets us reduce to $|\mathrm{Re}\,\tau| \leq 1$, and $\tau \to -1/\tau$ still lets us reduce to $|\tau| \geq 1$. The group action still identifies the left and right hand halves of the boundary of the fundamental domain, and the result is a Riemann surface of genus zero, now with two points omitted. The missing points are the Neveu-Schwarz cusp at $\tau = i\infty$ and the Ramond cusp at $\tau = \pm 1$ (those two points are equivalent under $\tau \to \tau + 2$). The Ramond cusp is missing in the quotient of the upper half plane by Γ_θ because the points $\tau = \pm 1$ are not in the upper half plane, but on its boundary.

So in terms of the function F, the behavior at the two cusps is determined by the low-lying spectrum in the NS and Ramond sectors, respectively.

We therefore call these the NS and R cusps. The natural uniformizing parameter at the NS cusp is $q^{1/2} = \exp(i\pi\tau)$, where τ is the argument of the function F in (3.25). The right parameter is $\exp(i\pi\tau)$ since the symmetry is $\tau \to \tau + 2$. At the R cusp, the natural parameter is $q = \exp(2\pi i\tau)$, where now τ is the argument of H.

To get a clear picture of the nature of a holomorphic function F on $\mathcal{H}/\Gamma_\theta$, it is necessary to describe the behavior near the cusps, or equivalently to describe what happens when one tries to extend F to a function on the compactification of $\mathcal{H}/\Gamma_\theta$ that is obtained by adding the cusps. This compactification is a space $\mathcal{Y} \cong \mathbb{CP}^1$.

We can explicitly describe a function K that parametrizes $\mathcal{H}/\Gamma_\theta$ and has a pole only at the NS cusp. This can be done in several ways. One formula is

$$K(\tau) = \frac{\Delta(\tau)^2}{\Delta(2\tau)\Delta(\tau/2)} - 24, \tag{3.34}$$

where $\Delta = q\prod_{n=1}^{24}(1-q^n)^{24}$ is the discriminant, a modular form of weight 24. In (3.34), we have subtracted the constant 24 so that the expansion of $K(\tau)$ in powers of $q^{1/2}$ has no constant term:

$$\begin{aligned} K(\tau) \ &= q^{-1/2} + 276q^{1/2} + 2048q + 11202q^{3/2} + 49152q^2 + 184024q^{5/2} + 614400q^3 \\ &\quad + 1881471q^{7/2} + 5373952q^4 + 14478180q^{9/2} + \dots . \end{aligned} \tag{3.35}$$

This is analogous to the definition of the J-function without a constant term. Another formula for K is

$$K = \frac{q^{-1/2}}{2}\left(\prod_{n=1}^{\infty}(1+q^{n-1/2})^{24} + \prod_{n=1}^{\infty}(1-q^{n-1/2})^{24}\right) + 2048q\prod_{n=1}^{\infty}(1+q^n)^{24}. \tag{3.36}$$

The product formula in (3.34), since it converges for all $|q|<1$, shows that K is non-singular as a function on $\mathcal{H}$. As for the behavior at the cusps, either formula shows that K has a simple pole at the NS cusp, that is, it behaves for $q\to 0$ as $q^{-1/2}$. K is regular at the Ramond cusp; in fact

$$K(\tau = 1) = -24. \tag{3.37}$$

This statement is equivalent to the statement that $K+24 = \Delta(\tau)^2/\Delta(2\tau)\Delta(\tau/2)$ vanishes at $\tau = 1$. Indeed, that function has a pole at the NS cusp, so it must have a zero somewhere. Its representation as a convergent infinite product shows that it is nonzero for $0<|q|<1$, so the zero is at the Ramond cusp. We give another explanation of (3.37) in section 3.3 in analyzing the $k^* = 1$ model. The fact that the holomorphic function K on $\mathcal{Y}$ has only one pole, which is of first order, gives another way to prove that $\mathcal{Y}$ is of genus zero.

Now let us consider the Neveu-Schwarz partition function F of a holomorphic SCFT. Any Γ_θ-invariant function F on $\mathcal{H}$ can be written as a function of K. The function F arising in a holomorphic SCFT is actually *polynomial* in K. Indeed, since the definition of F as $\mathrm{Tr}\, q^{L_0}$ is convergent for $0<|q|<1$, F is regular as a function on $\mathcal{H}$. So the only poles of F are at cusps, but the formula (3.31), which reflects the fact that the ground state energy in the Ramond sector is zero, says that F has no pole at the Ramond cusp.

So the only pole of F is at the Neveu-Schwarz cusp, that is at $K = \infty$. Consequently, in any holomorphic SCFT, the Neveu-Schwarz partition function F is a polynomial in K.

The degree of this polynomial is precisely k^*, since $F \sim q^{-k^*/2}$ for $q \to 0$. So

$$F = \sum_{r=0}^{k^*} f_r K^r. \tag{3.38}$$

Thus, F depends on $k^* + 1$ coefficients.

Now, for the case of an SCFT that is dual to three-dimensional supergravity, what would we expect F to be? To follow the same logic as in the bosonic case, we would argue that since the minimum energy of a BTZ black hole is $L_0 = 0$, and the entropy vanishes at $L_0 = 0$, primary states corresponding to black holes should be absent for $L_0 \leq 0$. Equivalently, primary fields other than the identity should be absent for dimension less than $(k^* + 1)/2$. Either statement would mean that up to terms of order $q^{1/2}$, F would coincide with the naive function

$$F_0(k^*) = q^{-k^*/2} \prod_{n=2}^{\infty} \frac{1 + q^{n-1/2}}{1 - q^n} \tag{3.39}$$

that counts superconformal descendants of the identity.

For each positive integer k^*, there is a uniquely determined function F_{k^*} that is a polynomial in K and coincides with $F_0(k^*)$ up to order $q^{1/2}$. This is a natural analog of the partition function Z_k that we defined for an extremal CFT without supersymmetry.

However, the same logic that makes us think that there should be no NS primaries of $L_0 \leq 0$ also would make us believe that there should be no Ramond primaries of $L_0 = 0$. (There are automatically no Ramond states of $L_0 < 0$, since $L_0 = \mathcal{G}_0^2$.) In either the NS or Ramond sector, the classical black hole entropy vanishes for $L_0 \leq 0$.

The number of Ramond primaries of $L_0 = 0$ in a theory with Neveu-Schwarz partition function F_{k^*} is, according to (3.31), $h_0 = (-1)^{k^*} F_{k^*}(1)$. Let us call this number β_{k^*}. β_{k^*} is uniquely determined for each k^*. A practical way to determine it, using (3.37), is to write

$$\beta_{k^*} = (-1)^{k^*} F_{k^*}(1) = (-1)^{k^*} \sum_{r=0}^{k^*} f_r K(1)^r = (-1)^{k^*} \sum_{r=0}^{k^*} f_r (-24)^r. \tag{3.40}$$

So one first evaluates the coefficients f_r in (3.38) to ensure that there are no NS primary states with $L_0 \leq 0$, and then one evaluates the sum in (3.40).

The first ten values of β_{k^*} are given in Table 1. The first observation is that β_{k^*} never vanishes in this range. Consequently, it is impossible to assume that there are no primary operators other than the identity of dimension less than $(k^*+1)/2$. If we assume that there are no such primaries in the NS sector, then there are primaries of dimension $k^*/2$ in the Ramond sector.

k^*	β_{k^*}
1	24
2	24
3	95
4	1
5	143
6	1
7	262
8	-213
9	453
10	-261

Table 1. Values of the index β_{k^*} for $k^* = 1, \ldots, 10$.

The good news is that the numbers in the table are rather small compared to black hole multiplicities that arise in the classically allowed region $L_0 > 0$. So an optimistic view is to interpret the numbers in the table as quantum corrections. For example, at $k^* = 9$, where β takes the relatively large value 453, the multiplicity of the lowest mass classically allowed black hole, namely $L_0 = 1/2$ in the NS sector, turns out to be 135149371 if the partition function F_9 can be trusted. (As usual, we count only primaries, although this involves only a small correction.) The lowest classically allowed black hole in the Ramond sector, at $L_0 = 1$, has multiplicity 381161020987.

However, it may not seem logical to assume that there are no NS primaries at dimension $k^*/2$ and allow Ramond primaries of that dimension. Moreover, it really does not make sense to do this, since some values of β in the table are negative.

So we retreat from claiming that there are no NS primary fields of dimension less than $(k^*+1)/2$, and instead we consider the hypothesis that there are no such primaries

of dimension less than $k^*/2$. In this case, we are free to add an integer s to F_{k^*}. If we do so, the number of NS primaries of dimension $k^*/2$ becomes $f_0 = s$, and the number h_0 of Ramond primaries of that dimension changes by $(-1)^{k^*}s$. The combination

$$\beta_{k^*} = h_0 - (-1)^{k^*} f_0 \tag{3.41}$$

is unchanged, and so is still given by the numbers in the table.

Unfortunately, then, we have no way to determine the numbers f_0 and h_0 separately, only the "index" β. We also have little insight about the more conventional index χ defined in (3.23). Both of these quantities characterize the primary states of $L_0 = 0$.

As in the bosonic case, one can construct useful formulas for the functions F_{k^*} in terms of Hecke operators acting on F_1. However, we will omit this.

Optimistic Conjecture

We define an extremal SCFT to be one with no primary fields other than the identity of dimension less than $k^*/2$. Equivalently, the NS partition function is equal up to an additive integer to the function F_{k^*} that was defined above.

Generalizing what we have said about the bosonic case, the most optimistic conjecture we can propose is that an extremal SCFT exists and is unique for every positive integer k^*, and is dual to three-dimensional supergravity. The best evidence we can offer is that extremal SCFT's in this sense do exist for $k^* = 1, 2$ and there are results about uniqueness [31,35] at least for $k^* = 1$.

In addition, as we will see, the functions F_{k^*} with $k^* = 3, 4$ have interesting properties suggestive of the existence of actual theories.

These matters are discussed in section 3.3. In addition, in the appendix, we describe the functions F_{k^*} for $5 \leq k^* \leq 10$.

3.3. Extremal SCFT's With Small k^*

k^*=1

The first construction of an extremal SCFT was made by FLM at $k^* = 1$, that is $c = 12$. They considered eight free bosons X_i compactified using the E_8 root lattice, combined with eight free fermions ψ_i to achieve superconformal symmetry. This theory has NS primary fields of dimension 1/2, namely the ψ_i. To eliminate these fields, they considered a $\mathbb{Z}_2$ orbifold, dividing by the operation that acts as -1 on all X_i and ψ_i.

They conjectured that this construction gave the unique SCFT that has $c = 12$ and no NS primary of dimension $1/2$.

While the construction is simple, it has the drawback of not making manifest the global symmetry of the model. This was remedied much more recently by Duncan [35], who described the same model in another way. In this construction, one begins with 24 free fermions λ_i, $i = 1, \dots, 24$, forming again a system with $c = 12$. In quantization, one can require the λ_i to be either antiperiodic or periodic around the spatial direction, giving what we will call NS_0 and R_0 sectors. (We reserve the name NS and R for another construction that will appear shortly.)

With 24 fermions, the ground state energy in the NS_0 sector is $-1/2$, while in the R_0 sector, the ground state energy is $+1$. The minimum dimension of a spin field is the difference between these two energies, or $3/2$.

The model has an $O(24)$ symmetry that rotates the fields λ_i. The ground state in the R_0 sector is highly degenerate, because of zero modes of the fields λ_i. It consists of $2^{12} = 4096$ states, transforming in the spinor representation of $O(24)$ (or rather its double cover).

The spin fields of lowest dimension are therefore 4096 fields W_α of dimension $3/2$ transforming as spinors. Dimension $3/2$ is the correct dimension for a supercurrent, but a generic linear combination $W = \sum_\alpha \epsilon^\alpha W_\alpha$, with c-number coefficients ϵ^α, does not generate a superconformal algebra. Schematically, and without worrying about the precise coefficients, the operator product $W \cdot W$ gives

$$W(x)W(0) \sim \frac{\overline{\epsilon}\epsilon}{x^3} + \frac{\overline{\epsilon}\epsilon\, T}{x} + \frac{\overline{\epsilon}\Gamma^{ij}\epsilon\, \lambda_i\lambda_j}{x^2} + \frac{\overline{\epsilon}\Gamma^{ijkl}\epsilon\, \lambda_i\lambda_j\lambda_k\lambda_l}{x} + \text{regular}. \tag{3.42}$$

(Here T is the stress tensor and Γ_i are the gamma matrices of $O(24)$.) The condition for W to generate a superconformal algebra is that the last two terms should be absent. This is equivalent to

$$\overline{\epsilon}\Gamma^{ij}\epsilon = \overline{\epsilon}\Gamma^{ijkl}\epsilon = 0. \tag{3.43}$$

It is shown in [35] that a spinor ϵ obeying these conditions exists and, if normalized to have $\overline{\epsilon}\epsilon = 1$, is unique up to an $O(24)$ transformation. Another important fact is that any such ϵ has one definite $SO(24)$ chirality or the other.

Any choice of a solution of (3.43) turns the theory into an $\mathcal{N} = 1$ superconformal field theory, with W as the supercurrent. However, as ϵ is not $O(24)$-invariant, the theory regarded as an $\mathcal{N} = 1$ SCFT does not have $O(24)$ symmetry. Rather, the choice of ϵ breaks

$O(24)$ symmetry to a group that is known as the Conway group Co_0. It is the symmetry group of the Leech lattice, and is a double cover of a sporadic finite group Co_1. Thus, as a superconformal field theory, this model has symmetry Co_0, not $O(24)$.

With this choice of superconformal algebra, we need to understand what are the Neveu-Schwarz and Ramond vertex operators. A Ramond vertex operator $\mathcal{O}$ has a square root singularity in the presence of the supercurrent W:

$$W(x)\mathcal{O}(x') \sim \frac{\mathcal{O}'}{(x-x')^{n-1/2}} \tag{3.44}$$

for some integer n and some operator $\mathcal{O}'$. With W understood as a spin operator with respect to the original fermions λ, those fermions have precisely this property. So they are Ramond fields. This enables us to analyze all of the states in the original NS_0 sector. The ground state corresponds to the identity operator, whose OPE with W of course has no branch cut, so it is an NS rather than Ramond operator. The operators in the NS_0 sector that have no branch cut with W are those that are products of an even number of λ's and their derivatives. The partition function that counts the corresponding states is

$$\frac{q^{-1/2}}{2}\left(\prod_{n=1}^{\infty}(1+q^{n-1/2})^{24}+\prod_{n=1}^{\infty}(1-q^{n-1/2})^{24}\right). \tag{3.45}$$

We recognize this as part of the formula (3.36) for the function K. In the R_0 sector, of the 4096 ground states, half have one chirality or fermion number and half have the other. Any excitation at all can be combined with a ground state of properly chosen chirality to get an operator that either does or does not have a branch cut with W, as desired. So for each L_0 eigenvalue in the R_0 sector, precisely half the states contribute to the NS sector and half to the R sector. The contribution of R_0 states to the NS sector is therefore

$$2048q\prod_{n=1}^{\infty}(1+q^n)^{24}. \tag{3.46}$$

Adding up (3.45) and (3.46), we see that the total partition function F_1 of the NS sector in this model is precisely what we have called K:

$$F_1 = K = \frac{q^{-1/2}}{2}\left(\prod_{n=1}^{\infty}(1+q^{n-1/2})^{24}+\prod_{n=1}^{\infty}(1-q^{n-1/2})^{24}\right)+2048q\prod_{n=1}^{\infty}(1+q^n)^{24}. \tag{3.47}$$

We can similarly compute the Ramond partition function H_1 of this model. The contribution of the NS_0 sector is obtained from (3.45) by changing a sign so as to project

onto states of odd fermion number, rather than even fermion number. And the contribution of the R_0 sector is the same as (3.46). So

$$\begin{aligned} H_1 &= \frac{q^{-1/2}}{2}\left(\prod_{n=1}^{\infty}(1+q^{n-1/2})^{24} - \prod_{n=1}^{\infty}(1-q^{n-1/2})^{24}\right) + 2048q\prod_{n=1}^{\infty}(1+q^n)^{24} \\ &= 24 + 4096q + 98304q^2 + 1228800q^3 + 10747904q^4 + \dots . \end{aligned} \tag{3.48}$$

Except for states of $L_0 = 0$, the global supercharge $\mathcal{G}_0$ of the Ramond sector exchanges the part of the Ramond sector coming from NS_0 with the part coming from R_0. This implies that we can alternatively write

$$H_1 = 24 + 4096q\prod_{n=1}^{\infty}(1+q^n)^{24}, \tag{3.49}$$

where we have removed the NS_0 contribution except for the ground states, and doubled the R_0 contribution. The equivalence of these formulas is not very obvious.

A consequence of the above formulas is that h_0, the number of Ramond states of $L_0 = 0$, is equal to 24. This gives another explanation for (3.37). Another consequence of (3.49) and (3.47) is that for $L_0 > 0$, the coefficients h_n in the q-expansion of H_1 are precisely twice the corresponding coefficients f_n in the expansion of F_1:

$$h_n = 2f_n. \tag{3.50}$$

(Here h_n is defined for integer n, while f_n is defined for integer or half-integer n, so this formula involves all coefficients h_n but only half of the f_n.) This is highly exceptional for SCFT's and reflects the particular way that this one was constructed. The relation (3.50) is equivalent to

$$0 = F_1(\tau) + F_1(\tau+1) - H_1(\tau) = F_1(\tau) - G_1(\tau+1) - H_1(\tau). \tag{3.51}$$

Although disguised in our way of presenting it, this is a standard relation in string theory. In the FLM construction of the model via 8 free bosons and 8 free fermions, (3.51) amounts to the statement of Gliozzi, Olive, and Scherk [52] that the Ramond-Neveu-Schwarz model has equal partition function in the Ramond and Neveu-Schwarz sectors, because of space-time supersymmetry.

The factor of 2 in (3.50) is consistent with asymptotic equality between the numbers of NS and Ramond black holes because it only involves half of the coefficients of F_1. H_1 has only half as many coefficients as F_1, but they are twice as big.

As we have explained, the NS partition function F_{k^*} of an extremal SCFT with any k^* is a polynomial in K or equivalently in F_1,

$$F_{k^*} = f(F_1), \tag{3.52}$$

for some polynomial f. From (3.29), it follows that the Ramond partition function H_{k^*} of an extremal SCFT can be obtained from H_1 using the same polynomial:

$$H_{k^*} = (-1)^{k^*} f(-H_1). \tag{3.53}$$

A Model With $k^* = 2$

What in our language is an extremal SCFT of $k^* = 2$ was constructed [26] by Dixon, Ginsparg, and Harvey (DGH), relatively soon after the work of FLM, by twisting the FLM monster construction in a somewhat similar fashion. Note that an extremal SCFT of $k^* = 2$ has $c = 24$, like the FLM monster theory.

We recall that the starting point of the FLM construction is 24 free bosons X_i that are compactified via the Leech lattice. Then one performs a $\mathbb{Z}_2$ orbifold, dividing by the symmetry $X_i \to -X_i$. To construct an orbifold, one first constructs untwisted and twisted sectors (by quantizing fields $X_i(\sigma)$ that are assumed to be periodic or antiperiodic functions of σ) and then one projects both sectors onto their $\mathbb{Z}_2$-invariant subspaces.

DGH observed that the ground state energy in the untwisted sector is -1, while the ground state energy in the twisted sector is $+1/2$. So a twist operator of lowest energy has dimension $3/2$, the right dimension for a supercurrent. They went on to show that it is possible to pick a twist operator $\mathcal{S}$ that generates a superconformal algebra.

In the FLM construction, the field $\mathcal{S}$ is not present, since it is odd rather than even under $\mathbb{Z}_2$, and is projected out when one forms the orbifold theory. However, DGH showed that it is possible to obtain an $\mathcal{N} = 1$ SCFT by modifying the usual orbifold projection. They defined the NS sector to consist of fields that do not have a cut in their OPE with $\mathcal{S}$, while the Ramond sector consists of fields that do have such a cut. Concretely, the NS sector consists of the $\mathbb{Z}_2$-even part of the untwisted sector of the orbifold plus the part of the twisted sector that transforms under $\mathbb{Z}_2$ as does $\mathcal{S}$ (so that $\mathcal{S}$, in particular, is an NS field), while the Ramond sector consists of the $\mathbb{Z}_2$-odd part of the untwisted sector plus the $\mathbb{Z}_2$-even part of the twisted sector.

From this information, one can of course compute the NS and Ramond partition functions F_2 and H_2. From our point of view, of course, F_2 is a polynomial in $F_1 = K$,

chosen so that $F_2 = q^{-1} + \mathcal{O}(q^{1/2})$. This gives $F_2 = K^2 - 552$, and similarly, in view of (3.53), $H_2 = H_1^2 - 552$. From this, we get

$$\begin{aligned} F_2 &= q^{-1} + 4096q^{1/2} + 98580q + 1228800q^{3/2} + 10745856q^2 + 74244096q^{5/2} + 432155586q^3 + \dots \\ H_2 &= 24 + 196608q + 21495808q^2 + 864288768q^3 + \dots . \end{aligned} \tag{3.54}$$

The number of states of $L_0 = 0$ is 24 in the Ramond sector and 0 in the NS sector.

The global symmetry of the DGH model is closely related to the symmetry group Co_0 of the Leech lattice (it is an extension of this by a finite abelian group that involves the momenta of the Leech lattice). Since the $k^* = 1$ model also has Co_0 symmetry, one might optimistically conjecture that this group is relevant to supergravity at all k^*. However, results below about $k^* = 3$ indicate that this is not the case.

Partition Function With $k^* = 3$

Unfortunately, this is the last case for which a suitable SCFT is known, but we can of course determine the appropriate NS and Ramond partition functions for all k^*, at least modulo an additive integer. We consider $k^* = 3, 4$ here because they turn out to have unusual properties, and relegate the further cases $k^* = 5, 6, \dots, 10$ to an appendix.

For $k^* = 3$, we have

$$F_3 = K^3 - 828K - 6143. \tag{3.55}$$

This leads to

$$\begin{aligned} F_3 &= q^{-3/2} + 1 + 33606q^{1/2} + 1843200q + 43434816q^{3/2} + 648216576q^2 \\ &\quad + 7171304841q^{5/2} + 63903727616q^3 + \dots \\ H_3 &= 95 + 3686400q + 1296433152q^2 + 127807455232q^3 + \dots . \end{aligned} \tag{3.56}$$

Here one finds an identity just like (3.50): the expansion coefficients h_n of H_3 are related for $n > 0$ to analogous coefficients f_n of F_3 by

$$h_n = 2f_n. \tag{3.57}$$

This expresses the fact that

$$F_1^3(\tau) + F_1^3(\tau + 1) = H_1^3(\tau) - 1536, \tag{3.58}$$

along with the similar linear relation (3.51).

It seems that the identity $h_n = 2f_n$, $n > 0$, does not hold for any values of k^* except 1 and 3. In general, the differences $h_n - 2f_n$ are relatively small, reflecting the fact that asymptotically an NS or Ramond black hole has the same entropy, but they are not zero. The fact that these degeneracies are actually equal for $k^* = 3$ may be a clue to the construction of this model.

One important point about the $k^* = 3$ model is that it seems very unlikely to have Co_0 or Co_1 symmetry. Indeed the leading coefficient 95 of H_3 is not in any economical way the dimension of a representation of Co_0. (The dimensions of the first few irreducible representations of this group are 1, 24, 276, 299, and 1771; for Co_1, one must omit the number 24 from this list.) Adding an integer n to F_3 and therefore subtracting n from H_3, does not help; we cannot pick n so that n and $95 - n$ are both dimensions of Co_0 representations in an economical fashion. So it appears that, despite the tempting evidence from the cases $k^* = 1, 2$, the group Co_0 is probably not a general symmetry of three-dimensional supergravity.

Monster Symmetry For $k^* = 4$?

However, in the next case, $k^* = 4$, there may well be a much larger discrete symmetry group, namely the monster group $\mathbb{M}$.

For $k^* = 4$, we have

$$F_4 = K^4 - 1104K^2 - 8191K + 107545. \tag{3.59}$$

This leads to the NS and Ramond partition functions

$$\begin{aligned} F_4 &= q^{-2} + q^{-1/2} + 1 + 196884q^{1/2} + 21493760q + 864299970q^{3/2} + 20246053140q^2 \\ &\quad + 333202640600q^{5/2} + 4252023300096q^3 + 44656994071935q^{7/2} + 401490908149760q^4 + \dots \\ H_4 &= 1 + 42987520q + 40491712512q^2 + 8504046600192q^3 + 802981773312000q^4 + \dots . \end{aligned} \tag{3.60}$$

The first non-trivial coefficient in F_4 is the famous number $196884 = 196883 + 1$ whose appearance in the J function gave the original hint of a connection between $\mathbb{M}$ and modular functions and ultimately conformal field theory. Moreover, the leading coefficient in H_4 is 1, which is compatible with any assumption about an automorphism group, since it is the dimension of an irreducible representation, namely the trivial one. The contrasts with what we found for $k^* = 3$, where the leading coefficient, namely 95, makes it difficult to postulate a large discrete symmetry.

To explore further the hypothesis of $\mathbb{M}$ symmetry, we attempt to express the coefficients in F_4 and H_4 in terms of dimensions of representations of $\mathbb{M}$. The dimensions $d_1, d_2, \ldots, d_{12}$ of the first 12 monster representations, which we call R_i, $i = 1, \ldots, 12$, are given in the table.

d_1	1
d_2	196883
d_3	21296876
d_4	842609326
d_5	18538750076
d_6	19360062527
d_7	293553734298
d_8	3879214937598
d_9	36173193327999
d_{10}	125510727015275
d_{11}	190292345709543
d_{12}	222879856734249

Table 2. Presented here from [53] are the dimensions d_i of the ith irreducible representation of the monster group $\mathbb{M}$, for $i = 1, \ldots, 12$. We denote as R_i the representation of dimension d_i.

A little experimentation soon shows that the nontrivial coefficients in the NS partition function F_4 can be nicely written in terms of the d_i. The first 8 nontrivial coefficients $f_{1/2}, f_1, \ldots, f_4$ are

$$\begin{aligned}
196884 &= d_1 + d_2 \\
21493760 &= d_1 + d_2 + d_3 \\
864299970 &= 2d_1 + 2d_2 + d_3 + d_4 \\
20246053140 &= 3d_1 + 4d_2 + 2d_3 + d_4 + d_6 \\
333202640600 &= 4d_1 + 5d_2 + 3d_3 + 2d_4 + d_5 + d_6 + d_7 \\
4252023300096 &= 5d_1 + 7d_2 + 4d_3 + 4d_4 + 2d_5 + 2d_6 + d_7 + d_8 \\
44656994071935 &= 7d_1 + 11d_2 + 7d_3 + 6d_4 + 3d_5 + 4d_6 + 2d_7 + 2d_8 + d_9 \\
401490908149760 &= 10d_1 + 16d_2 + 12d_3 + 9d_4 + 5d_5 + 7d_6 + 4d_7 + 4d_8 + d_9 + d_{10} + d_{12}.
\end{aligned} \tag{3.61}$$

The coefficients on the right hand side are rather small given the numbers involved, and as we will explain shortly, they do not grow much faster than is required by superconformal symmetry. The first three numbers here equal the first three non-trivial coefficients of the J function, but afterwards the two series diverge.

The coefficients h_n in the q-expansion of H_4 can similarly be expressed in terms of dimensions of monster representations. In doing so, however, the following is a useful shortcut and also possibly a clue to constructing the model. First of all, in contrast to the case of $k^* = 1, 3$, it is not true that $h_n - 2f_n = 0$ for $k^* = 4$ and all $n > 0$. But some of these differences vanish, and they are all relatively small. As we have already expressed the f_n for small n in terms of monster dimensions in (3.61), it suffices now to find similar expressions for $h_n - 2f_n$. Actually, as the h_n are all even for $n > 0$, we prefer to divide by 2, and we find

$$\begin{aligned}
\frac{h_1}{2} - f_1 &= 0 \\
\frac{h_2}{2} - f_2 &= -f_{1/2} \\
\frac{h_3}{2} - f_3 &= 0 \\
\frac{h_4}{2} - f_4 &= -f_1 \\
\frac{h_5}{2} - f_5 &= 0.
\end{aligned} \tag{3.62}$$

When this is combined with (3.61), we see that the $h_i/2$, $i \leq 5$ are linear combinations of the d_i with small positive integer coefficients. So the initial coefficients of the Ramond partition function are compatible with $\mathbb{M}$ symmetry, and in addition many of the differences $h_n/2 - f_n$ vanish.

T. Gannon has given a simple explanation of these results by showing that one can express F_4 and H_4 in terms of the J-function via $F_4(\tau) = J(2\tau) + J(\tau/2) + 1$, $H_4(\tau) = J(\tau/2) + J((\tau+1)/2) + 1$. Hence F_4 and H_4 inherit a relation to the monster from J. The fact that the formulas have such a direct explanation may lessen the case for a new SCFT with monster symmetry.

Given that the above formulas exist, they are not quite unique, because there are

some linear relations[17] among the d_i with small coefficients:

$$\begin{aligned} d_1 + d_4 + d_5 &= d_3 + d_6 \\ d_3 + d_8 + d_{12} &= d_2 + d_7 + d_9 + d_{11}. \end{aligned} \tag{3.63}$$

We have made some choices to ensure that the above formulas are compatible with superconformal symmetry, in the following sense. The first formula in (3.61) indicates that at $L_0 = 1/2$, the NS sector has 196883 primary states $|\rho_i\rangle$ transforming in the representation R_2. It follows that at $L_0 = 1$, there are descendants $\mathcal{G}_{-1/2}|\rho_i\rangle$ transforming in the same representation. This continues at higher levels; the number of copies of R_2 appearing as descendants of the primary states $|\rho_i\rangle$ at $L_0 = s/2$ is $1, 1, 1, 2, 3, 4, 5, 7$ for $s = 1, 2, \dots, 8$. If we count only primary states, we find, assuming that (3.61) is the right decomposition, that primary states in the representation R_2 occur at $L_0 = 1/2, 3/2, 2, 7/2, 4$, each time with multiplicity 1. Similar remarks apply to other representations. A special case is that, for the range of L_0 considered in (3.61), most of the states that transform in the trivial representation R_1 are actually descendants of the identity. If (3.61) is the right decomposition, then the first primary field that is $\mathbb{M}$-invariant is at $L_0 = 5/2$. If we were to rewrite (3.61) in terms of primary states only, the coefficients would be even smaller (but still nonnegative), strengthening the case for monster symmetry.

In this particular example, we cannot add an integer to F_4 without spoiling the hypothesis of monster symmetry. If there were an NS primary with $L_0 = 2$, it would have a descendant at $L_0 = 5/2$, and the number of primaries at $L_0 = 5/2$ would be less than 196883.

In the appendix, we present somewhat similar though less extensive evidence for baby monster symmetry at $k^* = 6$.

Sum Over Spin Structures

[17] These relations can be explained as follows. The symmetric part of $R_2 \otimes R_2$ decomposes as $\mathrm{Sym}^2 R_2 \cong R_1 \oplus R_2 \oplus R_4 \oplus R_5$, and the antisymmetric part decomposes as $\wedge^2 R_2 \cong R_3 \oplus R_6$. On the other hand, $\dim(\mathrm{Sym}^2 R_2) - \dim(\wedge^2 R_2) = \dim R_2 = d_2$. Taken together, these facts imply the first relation in (3.63), and the second follows similarly by considering the decomposition of $R_3 \otimes R_3$.

Given an SCFT with c an integer multiple[18] of 24, by summing over spin structures, one can make an ordinary bosonic CFT. The sum over spin structures projects both the NS and Ramond sectors onto bosonic states of integer dimension, as is appropriate for a bosonic CFT.

For c to be a multiple of 24, k^* must be even. It was already shown by DGH that the sum over spin structures, applied to their model which in our language is $k^* = 2$, gives the FLM model at $k = 1$. A natural question (raised by J. Duncan) is what this operation does at higher even values of k^*.

The sum over spin structures removes fermionic operators such as the supercurrent $\mathcal{S}$, of dimension 3/2. However, it leaves operators such as $\mathcal{S}\partial\mathcal{S}$ that are even in $\mathcal{S}$. This operator has dimension 4, and is a Virasoro primary (though, of course, it is a descendant in the $\mathcal{N} = 1$ super-Virasoro algebra). A bosonic extremal CFT of $k \geq 4$ should not have a primary of dimension 4. Hence, the "sum over spin structures" operation applied to an extremal SCFT of $k^* \geq 8$ does not give an extremal CFT. However, it does give an extremal CFT if applied to a theory of $k^* = 4$ or 6.

Hence, if extremal SCFT's exist with $k^* = 4, 6$, they can be used to generate extremal CFT's of $k = 2, 3$. It is interesting that $k^* = 4, 6$ are precisely the values at which extremal SCFT's may have monster or baby monster symmetry. (The baby monster group is the centralizer of an involution in the monster, and hence it is conceivable for an SCFT at $k^* = 6$ to have baby monster symmetry while the CFT at $k = 3$ has monster symmetry. The framework for this is described in [26].)

Part of what we have said can be understood in another way. Supergravity and gravity are different in the semiclassical limit, and one cannot be obtained from the other by merely summing over spin structures. So whatever is the SCFT dual of supergravity (even if our assumptions are too optimistic and it is not extremal), the sum over spin structures applied to this SCFT cannot give, for arbitrarily large k^*, the CFT that is dual to gravity.

[18] If c is an odd multiple of 12, the sum over spin structures is modular-invariant but projects out the NS ground state and does not give a CFT in the usual sense. This is actually important in superstring theory, where, with $c = 12$ in the light-cone treatment, the NS ground state is a "tachyon" and is removed in the sum over spin structures.

4. $k = 2$ Partition Function On A Hyperelliptic Riemann Surface

The most important question raised by this paper is certainly whether appropriate CFT's and SCFT's exist, beyond the few examples that are known.

In this section, we will perform a small computation that aims to give a hint that this is the case, at least for the bosonic theory with $k = 2$ and hence $c = 48$. We will show that the partition function of such a theory on a hyperelliptic Riemann surface of any genus can be determined in a unique and consistent way. This includes, for example, any Riemann surface of genus 2. The fact that we get a unique and consistent result in this situation gives some encouragement for believing that the $k = 2$ model does exist and is unique.

In this paper, we merely demonstrate the consistency of an algorithm for determining the partition function. Hopefully it will be possible in future work to get explicit formulas, at least for genus 2. Our method also works for $k = 1$, though here the ability in principle to determine the partition function comes as no surprise, since the model has been constructed explicitly [24] and in fact the genus 2 partition function has been computed [54] by a quite different method.

Perhaps it would help orient the reader to compare what we will do to another possible approach to determining the partition function. Genus 2 partition functions were determined in [54] for a variety of holomorphic CFT's with $c = 24$. The basic method was to express the partition function in terms of a Siegel modular form, which depends on only finitely many coefficients, and determine the coefficients by considering the behavior when the Riemann surface C degenerates. There are two types of degeneration (C can break up into two genus 1 curves joined at a point, or can reduce to a single genus 1 curve with two points glued together). The partition function can be determined from the behavior at just one degeneration, and there is a problem of consistency to show that one gets the same result either way. One could attempt to demonstrate the consistency by studying the appropriate Siegel modular forms, but we prefer to prove consistency by establishing the associativity of a certain operator product algebra. One advantage is that this method can be used to determine the partition function on a hyperelliptic Riemann surface of any genus. On the other hand, if one could overcome the technicalities in genus 2 (see [55] for one approach), one could possibly compute the genus 2 partition function of an extremal CFT for all k.

4.1. Twist Fields

A hyperelliptic Riemann surface C is a double cover of the complex plane, for example a double cover of the complex x-plane, which we will call C_0, described by an equation

$$y^2 = \prod_{i=1}^{2g+2} (x - e_i). \tag{4.1}$$

The $2:1$ cover $C \to C_0$ is branched at the points $e_1, \dots, e_{2g+2}$. C is smooth if the e_i are distinct, and has genus g if the number of branch points is precisely $2g+2$.

Our approach to determining the partition function on such a Riemann surface is based on an old idea [56-59]. The partition function of a conformal field theory $\mathcal{W}$ on the hyperelliptic Riemann surface C can be determined by computing, in a doubled theory, the correlation function of $2g+2$ copies of a "twist field" $\mathcal{E}$, inserted at the points $e_1, \dots, e_{2g-2}$ in C_0.

Consider any CFT $\mathcal{W}$ of central charge c. Away from branch points, the theory $\mathcal{W}$ on the double cover C looks locally like the theory $\mathcal{W} \times \mathcal{W}$ on C_0. Here we have one copy of $\mathcal{W}$ for each of the two branches of $C \to C_0$.

In going around a branch point, the two copies of $\mathcal{W}$ are exchanged. So a more complete description is to say that the theory on C_0 is $\mathrm{Sym}^2\mathcal{W}$, the symmetric product of two copies of $\mathcal{W}$. The symmetric product theory is an orbifold of the product theory $\mathcal{W} \times \mathcal{W}$ in which one divides by the $\mathbb{Z}_2$ symmetry that exchanges the two branches.

Let us describe what kind of operators the symmetric product theory has. First, there are "untwisted" operators. These are simply operators of the theory $\mathcal{W} \times \mathcal{W}$ that are invariant under the exchange of the two factors. For example, let T^+ and T^- be the stress tensors of the two factors. Any local operator constructed as a polynomial in T^+ and T^- and their derivatives and invariant under the exchange $T^+ \leftrightarrow T^-$ gives an operator in the symmetric product theory. Operators of this particular type generate a chiral algebra that we will call $\mathrm{Sym}^2\mathcal{V}$, where $\mathcal{V}$ denotes the chiral algebra generated by a single stress tensor.

In addition, there are "twisted sector" operators. These operators correspond (in the operator/state correspondence of CFT) to states obtained by quantizing the theory $\mathcal{W} \times \mathcal{W}$ on a circle S^1, in such a way that the two copies of $\mathcal{W}$ are exchanged in monodromy around the circle. These states are the states of a *single* copy of $\mathcal{W}$ on a circle of twice the circumference. The energies of the states in the twisted sector of the $\mathrm{Sym}^2\mathcal{W}$ theory are therefore precisely one-half the energies of the original $\mathcal{W}$ theory (in conformal field theory, doubling the circumference of the circle divides the Hamiltonian by two).

Operators related to states in the twisted sector are called twist fields. The dimension of a twist field is the same as the difference in energy between the corresponding twisted sector state and the untwisted ground state. The ground state energy in the untwisted sector is $-2 \cdot c/24$ (where the 2 comes from the two copies of $\mathcal{W}$) and the ground state energy in the twisted sector is $-(1/2) \cdot c/24$ (where the factor of $1/2$ was explained in the last paragraph). The difference is $d_{\mathcal{E}} = (c/24)(-1/2 - (-2)) = (3/2)c/24$, and this is the dimension of the twist field $\mathcal{E}$ of lowest energy. For our application, we are interested in the case that $c/24$ is an integer k, and hence $d_{\mathcal{E}} = 3k/2$.

Now we can explain how the partition function of the theory $\mathcal{W}$ on the hyperelliptic Riemann surface C defined in eqn. (4.1) can be interpreted as a genus zero correlation function in the theory $\mathrm{Sym}^2\mathcal{W}$. From the standpoint of the $\mathrm{Sym}^2\mathcal{W}$ theory on the x-plane, the role of the branch points is just to exchange the two copies of $\mathcal{W}$, via a twist field. On the double cover, there is no operator insertion at the branch points except the identity; the identity operator at a branch point corresponds in the downstairs description to the ground state in the twisted sector and hence to the twist operator $\mathcal{E}$ of lowest dimension. So the partition function on the double cover can be expressed in terms of the correlation function $\left\langle \mathcal{E}(e_1)\mathcal{E}(e_2)\dots\mathcal{E}(e_{2g+2})\right\rangle$ on the x-plane. (The precise statement involves the conformal anomaly; this is deferred to section 4.5.)

To compute this correlation function, we need to understand the chiral algebra that is obtained by extending the symmetric product $\mathrm{Sym}^2\mathcal{V}$ of two Virasoro algebras by the field $\mathcal{E}$. For this, we need to know what primary fields (of $\mathrm{Sym}^2\mathcal{V}$) appear in the operator product $\mathcal{E}\cdot\mathcal{E}$. To calculate the product $\mathcal{E}\cdot\mathcal{E}$, we need to look at the behavior when a pair of branch points approach each other. This situation is described locally by an equation

$$y^2 = (x-e)(x-e'), \tag{4.2}$$

and we are interested in the behavior for $e \to e'$. In that limit, the equation becomes $y^2 = (x-e)^2$, and the surface breaks up into two branches $C_\pm : y = \pm(x-e)$. Each branch is a copy of the complex plane. We can express the product $\mathcal{E}(e)\cdot\mathcal{E}(e')$ for $e \to e'$ as a sum of operators of the form $\mathcal{U}_+ \otimes \mathcal{U}_-$, where $\mathcal{U}_\pm$ is an operator in the original CFT $\mathcal{W}$ on the branch $C_\pm$. Moreover, the $\mathcal{U}_\pm$ are descendants of some Virasoro primary fields $\mathcal{O}_\pm$.

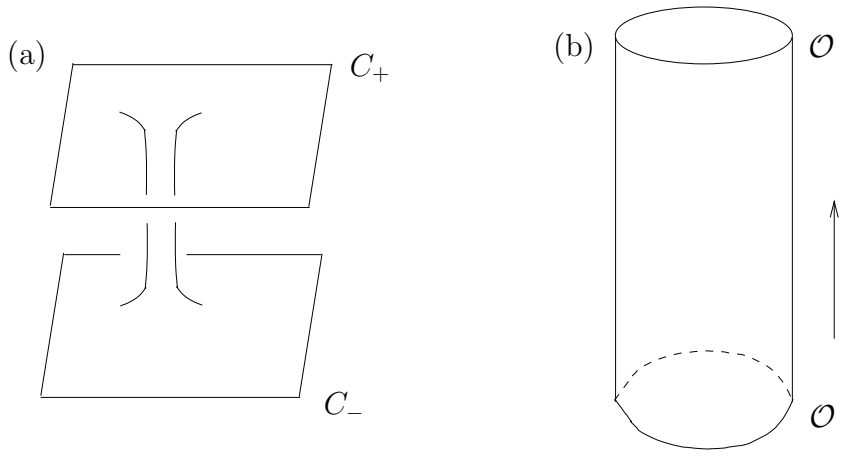


Fig. 5:

(a) The equation $y^2 = (x-e)(x-e')$ describes two branches connected by a tube. The tube collapses to a point as $e \to e'$. (b) Topologically, the tube is equivalent to a cylinder. Propagation along the tube multiplies any energy eigenstate by a c-number factor. If a primary field $\mathcal{O}$ flows in from the past, the same field flows out in the future. Alternatively, if we think of all fields as flowing out, then the two fields emerging at the two ends of the cylinder are conjugate.

The decomposition of $\mathcal{E}(e)\cdot\mathcal{E}(e')$ as a sum of descendants of operators $\mathcal{O}_+\otimes\mathcal{O}_-$ can be made in such a way that the operator $\mathcal{O}_+$ is conjugate to $\mathcal{O}_-$, meaning that the two-point function $\langle\mathcal{O}_+\mathcal{O}_-\rangle$ (in the original theory $\mathcal{W}$) is nonzero in genus zero. To see this, go back to the smooth double cover with $e \neq e'$. For e near e', the two branches are smoothly connected by a tube (fig. 5a). The surface thus has the topology of a cylinder (fig. 5b). Propagation along this cylinder is the "identity operator" on primary fields. So if a given primary state flows in at the bottom, it or flows out at the top. When we write $\mathcal{E}(e)\cdot\mathcal{E}(e')$ as a sum of descendants of operators $\mathcal{O}_+\otimes\mathcal{O}_-$, the convention is that both $\mathcal{O}_+$ and $\mathcal{O}_-$ are outgoing, so they are conjugate, rather than being equal. At any rate, the important conclusion for us is that $\mathcal{O}_+$ and $\mathcal{O}_-$ have the same dimension.

According to our interpretation of three-dimensional gravity, it is described by a conformal field theory with $c = 24k$ and no primary field other than the identity of dimension less than $k+1$. Hence, a primary operator $\mathcal{O}_+\otimes\mathcal{O}_-$ of the symmetric product theory, where $\mathcal{O}_+$ and $\mathcal{O}_-$ have the same dimension, is either the identity operator or has dimension at least $2(k+1)$.

To understand the chiral algebra or superalgebra[19] generated by $\mathcal{E}$ together with $\mathrm{Sym}^2\mathcal{V}$, we need to know all primary fields that appear with singular coefficients in the operator product $\mathcal{E}(e)\cdot\mathcal{E}(e')$. Since $\mathcal{E}$ has dimension $3k/2$, an operator that will contribute a singularity must have dimension less than $3k$ and hence at most $3k-1$. On the other hand, we have just seen that all operators appearing in the OPE are either descendants of the identity or have dimension at least $2(k+1)$.

For $k=1$ or 2, we have $3k-1<2(k+1)$. This leads to a drastic simplification, which is the reason that we will restrict ourselves here to $k=1,2$: singularities in the product $\mathcal{E}(e)\cdot\mathcal{E}(e')$ come only from the identity operator and its $\mathrm{Sym}^2\mathcal{V}$ descendants. Therefore, to understand correlation functions of $\mathcal{E}$, we need only understand the chiral algebra that is obtained by adding to $\mathrm{Sym}^2\mathcal{V}$ the primary field $\mathcal{E}$ of dimension $3k/2$ with the OPE schematically

$$\mathcal{E}\cdot\mathcal{E}\sim 1+\text{descendants}. \tag{4.3}$$

This is a chiral algebra that can be described in closed form and explicitly proved to obey the Jacobi identity. The only $\mathrm{Sym}^2\mathcal{V}$ primaries are 1 and $\mathcal{E}$. The only genus zero three point functions of primaries are $\langle 1\cdot 1\cdot 1\rangle$, which is trivial, and $\langle 1\cdot\mathcal{E}\cdot\mathcal{E}\rangle$, which is almost trivial in the sense that it reduces to a two-point function. To understand correlation functions of descendants, it suffices to understand the module for $\mathrm{Sym}^2\mathcal{V}$ consisting of $\mathcal{E}$ and its descendants. One way to construct this module explicitly is to observe that on the double cover of the x-plane, $\mathcal{E}$ simply corresponds to the identity operator and $\mathrm{Sym}^2\mathcal{V}$ to the ordinary Virasoro algebra; so this module for $\mathrm{Sym}^2\mathcal{V}$ can be deduced from the identity module for $\mathcal{V}$.

To get beyond the three-point function, the key step is to show that the chiral algebra obtained by adding $\mathcal{E}$ to $\mathrm{Sym}^2\mathcal{V}$ is consistent, that is, that the Jacobi identity is obeyed. If so, this chiral algebra determines all genus zero correlation functions of $\mathcal{E}$, and hence determines the partition function of our CFT on a general hyperelliptic Riemann surface.

In such a problem of trying to extend a known chiral algebra $\mathrm{Sym}^2\mathcal{V}$ by additional primary fields $\mathcal{E}$ (of integer or half-integer dimension, in which case one will get an extended chiral algebra or a superalgebra), the Jacobi identity is equivalent [60] to the statement that there exists a four-point function of the additional primary fields that has the appropriate singularities determined by the operator product expansion in all channels. (Such a

[19] As $\mathcal{E}$ has dimension $3k/2$, it is a fermionic operator if k is odd.

function automatically has the right symmetries.) In the case at hand, since the only primary field that we are trying to add is $\mathcal{E}$, the only non-trivial primary four-point function that we have to consider is $\langle \mathcal{E}(e_1)\mathcal{E}(e_2)\mathcal{E}(e_3)\mathcal{E}(e_4)\rangle$. In section 4.3, we explicitly construct this four-point function and show that it behaves correctly in all channels.

In the following sense, this result is not at all surprising. The four-point function $\langle \mathcal{E}(e_1)\mathcal{E}(e_2)\mathcal{E}(e_3)\mathcal{E}(e_4)\rangle$ is essentially the partition function of the underlying CFT on the hyperelliptic Riemann surface $y^2 = (x-e_1)(x-e_2)(x-e_3)(x-e_4)$. That surface has genus 1, and we already know from section 3 that for every k there is a unique, natural genus 1 partition function with the properties we want. In section 4.4, we show explicitly, for $k = 1, 2$, that the four-point function $\langle \mathcal{E}(e_1)\mathcal{E}(e_2)\mathcal{E}(e_3)\mathcal{E}(e_4)\rangle$ determined from the OPE's agrees with the genus 1 partition function as described in section 3.

4.2. The $\mathcal{E}\cdot\mathcal{E}$ Operator Product

As a first step in that direction, we will calculate the details of the $\mathcal{E}\cdot\mathcal{E}$ operator product. For this, we start with a double cover C of the x-plane branched at e and $e' = -e$, and so described by an equation $y^2 = x^2 - e^2$. If $u = x + y$, $v = x - y$, then the equation is $uv = e^2$. The two branches C_+ and C_- correspond respectively to $u \to \infty$, $v \to 0$ and $u \to 0$, $v \to \infty$.

The path integral over C gives a quantum state Ψ in the theory $\mathcal{W}\times\mathcal{W}$, that is, one copy of $\mathcal{W}$ for each branch. This state is invariant under exchange of the two branches by the symmetry $y \to -y$, so it is really a state in the symmetric product theory $\mathrm{Sym}^2\mathcal{W}$. We are really only interested in the part of Ψ proportional to the vacuum state and its descendants. As we have seen, this part suffices to describe the desired chiral algebra if $k \leq 2$.

We will determine Ψ by using the fact that certain elements in the product $\mathcal{V}\times\mathcal{V}$ of two Virasoro algebras annihilate Ψ. This is so because there are globally-defined holomorphic vector fields on C, of the form $V_n = 2^{-n}u^{n+1}d/du = -2^{-n}(e^2/v)^n v d/dv$. Let $\mathcal{S}$ be a contour on the surface C that wraps once around the "hole." If T is the stress tensor, the contour integral $\int_{\mathcal{S}} V_n T$ can be regarded, for any n, as an operator acting on the state Ψ. This operator is invariant under deformation of the contour. It can be deformed (fig. 6) to a contour $\mathcal{S}_+$ in the upper branch C_+ or a contour $\mathcal{S}_-$ in the lower branch C_-. So we have for all n

$$\left(\int_{\mathcal{S}_+} V_n T - \int_{\mathcal{S}_-} V_n T\right)\Psi = 0. \tag{4.4}$$

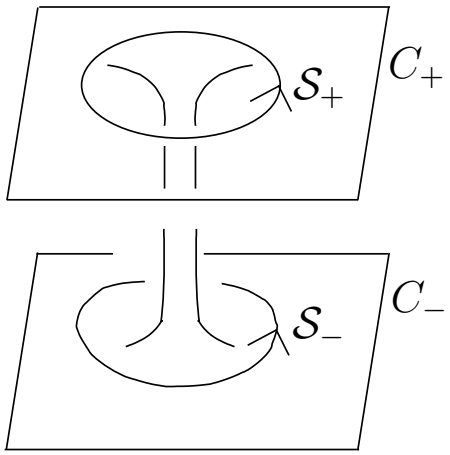


Fig. 6:

A contour $\mathcal{S}$ that wraps once around the hole can be deformed to the contour $\mathcal{S}_+$ in the upper branch or to the contour $\mathcal{S}_-$ in the lower branch.

We want to express the two contour integrals that appear here in terms of Virasoro generators on the two branches. To do this, we simply express V_n as a vector field on the branch C_+ or C_-, either of which we identify with the x-plane. For example, on the branch C_+, we write explicitly $y = \sqrt{1 - e^2/x^2} = 1 - e^2/2x^2 - e^4/8x^4 + \mathcal{O}(e^6)$. We have carried the expansion far enough to determine (for $k \leq 2$) all singular terms in the product $\mathcal{E}(e) \cdot \mathcal{E}(-e)$. So

$$V_n = x^{n+1} \left(1 - (n+2)\frac{e^2}{4x^2} + \frac{e^4}{x^4}\left(\frac{n^2+n-4}{32}\right) + \mathcal{O}(e^6)\right) \frac{d}{dx}. \tag{4.5}$$

This means, if we ignore the conformal anomaly for the moment, that $\int_{\mathcal{S}_+} V_n T$ corresponds, on the branch C_+, to the operator

$$Q_n^+ = L_n^+ - \frac{n+2}{4} e^2 L_{n-2}^+ + \left(\frac{n^2+n-4}{32}\right) e^4 L_{n-4}^+ + \mathcal{O}(e^6). \tag{4.6}$$

As a check, one can verify that $[Q_n^+, Q_m^+] = (n-m)Q_{n+m}^+$. Similarly, $\int_{\mathcal{S}_-} V_n T$ corresponds on the branch C_- to the operator

$$Q_n^- = \left(\frac{e^2}{4}\right)^n \left(L_{-n}^- - e^2\left(\frac{-n+2}{4}\right) L_{-n-2}^- + e^4\left(\frac{n^2-n-4}{32}\right) L_{-n-4}^- + \mathcal{O}(e^6)\right). \tag{4.7}$$

The state Ψ is determined for each value of e (up to multiplication by a complex scalar) by the condition that $\widehat{Q}_n\Psi = 0$, where $\widehat{Q}_n = Q_n^+ - Q_n^-$. However, because of the Virasoro anomaly some c-number terms must be added to the above formulas, reflecting the conformal anomaly in the mapping from u to x. There is no such c-number contribution to $\widehat{Q}_0$, since it cancels between the two branches. The c-numbers in $\widehat{Q}_n$ for other n can be conveniently determined by requiring that $[\widehat{Q}_n, \widehat{Q}_m] = (n-m)\widehat{Q}_{n+m}$. For our purposes, the only formulas we need are

$$\begin{aligned}
\widehat{Q}_0 &= \left(L_0^+ - \frac{e^2}{2}L_{-2}^+ - \frac{e^4}{8}L_{-4}^-\right) - \left(L_0^+ - \frac{e^2}{2}L_{-2}^- - \frac{e^4}{8}L_{-4}^-\right) + \dots \\
\widehat{Q}_1 &= \left(L_1^+ - \frac{3e^2}{4}L_{-1}^+ - \frac{e^4}{16}L_{-3}^+\right) - \frac{e^2}{4}\left(L_{-1}^- - \frac{e^2}{4}L_{-3}^-\right) + \dots \\
\widehat{Q}_2 &= L_2^+ - e^2L_0^+ - 3ke^2 + \frac{e^4}{16}L_{-2}^+ - \frac{e^4}{16}L_{-2}^- + \dots .
\end{aligned} \tag{4.8}$$

Terms of order e^6 have been omitted. The constant in $\widehat{Q}_2$ was obtained from $[\widehat{Q}_0, \widehat{Q}_2] = -2\widehat{Q}_2$.

By requiring that $\widehat{Q}_m\Psi = 0$ for $m = 0, 1, 2$, and that Ψ converges to the Fock vacuum $|\Omega\rangle$ for $e \to 0$, we now find Ψ to be

$$\begin{aligned}
\Psi(e) = \Bigg(1 &+ \frac{e^2}{4}(L_{-2}^+ + L_{-2}^-) + \frac{e^4}{32}(L_{-4}^+ + L_{-4}^-) \\
&+ \frac{e^4}{32}(L_{-2}^+ + L_{-2}^-)^2 + \frac{e^4}{192k}L_{-2}^+L_{-2}^- + \dots\Bigg)|\Omega\rangle.
\end{aligned} \tag{4.9}$$

We can immediately use this formula to determine the singular part of the $\mathcal{E}\cdot\mathcal{E}$ OPE. We normalize $\mathcal{E}$ so that the most singular term is $\mathcal{E}(x/2)\mathcal{E}(-x/2) \sim 1/x^{3k}$. Then, setting $e = x/2$, we get

$$\begin{aligned}
\mathcal{E}(x/2)\mathcal{E}(-x/2) \sim \frac{1}{x^{3k}}\Bigg(1 &+ \frac{x^2}{16}T + \frac{x^4}{2^{10}}\partial^2 T \\
&+ \frac{x^4}{2^9}T\star T + \frac{x^4}{3k\cdot 2^{10}}T^+T^-\Bigg) + \mathcal{O}(x^{-3k+6}).
\end{aligned} \tag{4.10}$$

All we have done is to write, on the right hand side, the operator that corresponds to the state $\Psi(e)/x^{3k}$, for $e = x/2$. (The factor of $1/x^{3k}$ must be supplied by hand, since in defining $\Psi(e)$, we just normalized it so that the coefficient of the Fock vacuum is 1.) Also, T^+ and T^- are the stress tensors on the branches C_+ and C_-, respectively. $T = T^+ + T^-$ is the diagonal or total stress tensor, and similarly $L_n = L_n^+ + L_n^-$. In addition, we use the fact that in the correspondence between operators and states, the states $L_{-2}^\pm|\Omega\rangle$ correspond to the operators $T^\pm$, while $L_{-4}^\pm|\Omega\rangle$ correspond to $\frac{1}{2}\partial^2 T^\pm$. Finally, we have written $T\star T$ for the operator corresponding to the state $L_{-2}^2|\Omega\rangle$. $T\star T(0)$ can be obtained as the term of order x^0 in the operator product $T(x)\cdot T(0)$.

4.3. Determining The Four-Point Function

Our goal is to use the explicit formula (4.10) to show that the chiral algebra obtained by adjoining $\mathcal{E}$ to $\mathrm{Sym}^2\mathcal{V}$ obeys the Jacobi identity. The Jacobi identity is equivalent to the existence of a four-point function for primary fields with the right symmetry properties and the right OPE singularities in all channels. It is enough to consider primary fields with respect to the $\mathrm{Sym}^2\mathcal{V}$ algebra, which we already know to exist. The only non-trivial case is the four-point function of $\mathcal{E}$.

For $k = 1$, the singular part of the $\mathcal{E} \cdot \mathcal{E}$ operator product only involves the diagonal stress tensor $T = T^+ + T^-$. This is a substantial simplification; the extended chiral superalgebra obtained by incorporating $\mathcal{E}$ can be viewed an extension of an ordinary Virasoro algebra $\mathcal{V}$, rather than a symmetric product $\mathrm{Sym}^2\mathcal{V}$. In fact, this algebra is a familiar one, an $\mathcal{N} = 1$ super Virasoro algebra:

$$\mathcal{E}(x/2)\mathcal{E}(-x/2) \sim \frac{1}{x^3}\left(1 + \frac{x^2}{16}T\right) + \mathcal{O}(x). \tag{4.11}$$

The stress tensor T has $c = 48$ (as it is the diagonal stress tensor in the symmetric product of two copies of a $c = 24$ theory). A superconformal field theory with this central charge can be constructed from 32 chiral multiplets consisting of bosonic and fermionic fields ϕ_i and ψ_i, where $\langle\phi_i(x)\phi_j(0)\rangle = -\delta_{ij}\ln(x)$, and $\langle\psi_i(x)\psi_j(0)\rangle = \delta_{ij}/x$, for $i, j = 1, \ldots, 32$. In terms of these fields, we can take

$$\mathcal{E} = \frac{i}{4\sqrt{2}}\sum_{k=1}^{32}\psi_k\partial\phi_k. \tag{4.12}$$

With this free field realization, all correlation functions of $\mathcal{E}$ can be explicitly computed. For example, the four-point function is

$$\begin{aligned}\langle\mathcal{E}(e_1)\mathcal{E}(e_2)\mathcal{E}(e_3)\mathcal{E}(e_4)\rangle = \prod_{i=1}^{4}(de_i)^{3/2}\Bigg(\Bigg(&\frac{1}{(e_1-e_2)^3(e_3-e_4)^3} + \text{cyclic}\Bigg)\\ &+\frac{3}{32}\frac{1}{\prod_{i<j}(e_i-e_j)}\Bigg),\end{aligned} \tag{4.13}$$

where additional terms obtained by cyclic permutations of e_2, e_3, e_4 are to be added in the first term on the right. We will eventually compare this formula to the known genus 1 partition function of the monster theory of $k = 1$.

In (4.13), we included a factor of $(de)^{3/2}$ for each operator $\mathcal{E}(e)$ to reflect the fact that $\mathcal{E}$ has dimension 3/2. Thus the correlation "function" $\big\langle \mathcal{E}(e_1)\mathcal{E}(e_2)\mathcal{E}(e_3)\mathcal{E}(e_4)\big\rangle$ is most naturally understood as a 3/2-differential in each variable. This factor is often omitted in writing such formulas, and we will do so as an abbreviation except when the factor is important.

For $k = 2$, the singular part of $\mathcal{E}\cdot\mathcal{E}$ cannot be expressed in terms of the diagonal stress tensor T only, and there is no way to avoid using $\mathrm{Sym}^2\mathcal{V}$. Correspondingly, the chiral algebra for $k = 2$ seems to be unfamiliar, and it is considerably harder to get the analog of (4.13) for $k = 2$. Since we do not have an explicit realization of the algebra analogous to (4.12), we cannot compute the four-point function directly, and instead we will determine it by requiring the appropriate OPE singularities.

The function $\big\langle \mathcal{E}(e_1)\mathcal{E}(e_2)\mathcal{E}(e_3)\mathcal{E}(e_4)\big\rangle$, with $e_1,\dots,e_4 \in C_0 \cong \mathbb{CP}^1$, must be symmetric in all arguments and must have the appropriate OPE singularities in all channels. In addition, it is highly constrained by invariance under the action of $SL(2,\mathbb{C})$ on C_0. The most general "function," actually a cubic differential in each variable, that is symmetric, $SL(2,\mathbb{C})$-invariant, and consistent with at least the leading OPE singularity, is

$$\begin{aligned}\big\langle \mathcal{E}(e_1)\mathcal{E}(e_2)\mathcal{E}(e_3)\mathcal{E}(e_4)\big\rangle = \prod_{i=1}^4 (de_i)^3 \Bigg(&\left(\frac{1}{(e_1-e_2)^6(e_3-e_4)^6} + \text{cyclic}\right) \\ &+ A\left(\frac{1}{(e_1-e_2)^4(e_3-e_4)^4}\prod_{i=1,2,\,j=3,4}\frac{1}{e_i-e_j} + \text{cyclic}\right) \\ &+ B\frac{1}{\prod_{i<j}(e_i-e_j)^2}\cdot\Bigg)\,.\end{aligned} \tag{4.14}$$

Cyclic permutations of e_2, e_3, e_4 are to be added as shown, and the coefficients A and B must be determined to ensure that the OPE singularities are correct.

After a detailed calculation, one finds that the right values are

$$A = \frac{3}{16},\ B = \frac{12k+1}{2^{16}} \tag{4.15}$$

where of course $k = 2$. To get these formulas, one may for instance set $e_1 = -e_2 = e$, and use (4.10) to determine the singular behavior for $e \to 0$. In this way, one expresses the singular behavior of the four-point function in terms of three-point functions $\big\langle \mathcal{X}(0)\mathcal{E}(e_3)\mathcal{E}(e_4)\big\rangle$, where $\mathcal{X}$ is one of the descendants of the identity that appear in (4.10).

The coefficient A appears in the coefficient of $1/e^4$ for $e \to 0$, and in determining it, the only important choices of $\mathcal{X}$ are 1 and T. The three-point functions that we need are

$$\begin{aligned} \big\langle 1 \cdot \mathcal{E}(z)\mathcal{E}(w)\big\rangle &= \frac{1}{(z-w)^6} \\ \big\langle T(x)\mathcal{E}(z)\mathcal{E}(w)\big\rangle &= \frac{3}{(z-x)^2(w-x)^2(z-w)^4}. \end{aligned} \tag{4.16}$$

The first is immediate from the way we have normalized $\mathcal{E}$. The second, since the three fields involved are quasi-primary fields (they transform as primaries under $SL(2,\mathbb{C})$) is determined by $SL(2,\mathbb{C})$ invariance up to a constant multiple. The constant can be determined using the $T \cdot \mathcal{E}$ OPE (which reflects the fact that $\mathcal{E}$ is a primary of dimension $3 = 3k/2$) or the $\mathcal{E} \cdot \mathcal{E}$ OPE.

To determine B, we need three more three-point functions, namely

$$\begin{aligned} \big\langle \partial^2 T(0)\mathcal{E}(z)\mathcal{E}(w)\big\rangle &= \left(\frac{1}{z^2}+\frac{1}{w^2}\right)\frac{18}{z^2w^2(z-w)^4} + \frac{24}{z^3w^3(z-w)^4} \\ \big\langle T\star T(0)\mathcal{E}(z)\mathcal{E}(w)\big\rangle &= \left(\frac{5}{z^2}+\frac{5}{w^2}-\frac{4}{zw}\right)\frac{3}{z^2w^2(z-w)^4} \\ \big\langle T^+T^-(0)\mathcal{E}(z)\mathcal{E}(w)\big\rangle &= \left(\frac{9k^2}{16}+\frac{3k}{64}\right)\frac{1}{z^4w^4(z-w)^2}. \end{aligned} \tag{4.17}$$

The first of these results is obtained simply by differentiating the second formula in (4.16).

To get the second, we use the fact that

$$T\star T(0) = \mathrm{Res}_{x=0}\frac{dx\,T(x)}{x}T(0). \tag{4.18}$$

Now consider the expression

$$\left\langle \frac{dx\,T(x)}{x}T(0)\mathcal{E}(z)\mathcal{E}(w)\right\rangle. \tag{4.19}$$

We view this as a differential form on the complex x-plane, with z and w kept fixed. As such, it has poles precisely at $x = 0, z, w$. The residue at $x = 0$ is the desired three-point function $\big\langle T\star T(0)\mathcal{E}(z)\mathcal{E}(w)\big\rangle$. The residues at at $x = z, w$ can be computed in terms of the correlator $\big\langle T(0)\mathcal{E}(z)\mathcal{E}(w)\big\rangle$ by using the fact that $\mathcal{E}$ is a primary of dimension 3, which determines the singular behavior of the product $T\cdot\mathcal{E}$. The sum of the residues must vanish, and this gives our result.

Finally, to get the last formula in (4.17), we note first that since the operators T^+T^- and $\mathcal{E}$ are quasiprimary fields of dimension 4 and 3, we have

$$\left\langle T^+T^-(x)\mathcal{E}(z)\mathcal{E}(w)\right\rangle = \frac{\theta}{(x-z)^4(x-w)^4(z-w)^2}, \tag{4.20}$$

for some constant θ. We have

$$\lim_{w\to\infty}\frac{\left\langle T^+T^-(x)\mathcal{E}(0)\mathcal{E}(w)\right\rangle}{\left\langle \mathcal{E}(0)\mathcal{E}(w)\right\rangle} = \frac{\theta}{x^4}. \tag{4.21}$$

This is supposed to be computed in the theory $\mathrm{Sym}^2\mathcal{W}$ on the x-plane, with insertions of $\mathcal{E}$ at $0, w$. This is equivalent to a computation in the theory $\mathcal{W}$ on a double cover of the x-plane branched at $0, w$. In the limit $w \to \infty$, the double cover is described by the equation $y^2 = x$, and it is convenient to transfer the computation to the y-plane.

We will now write the stress tensor explicitly as T_{xx} or T_{yy} depending on which local parameter is used in defining it. Allowing for the conformal anomaly, the relation between the stress tensor T_{xx} on the x-plane and the corresponding stress tensor T_{yy} on the y-plane is

$$\left(\frac{\partial y}{\partial x}\right)^2 T_{yy} = T_{xx} - \frac{c}{12}\{y,x\}, \tag{4.22}$$

where $\{y,x\}$ denotes the Schwarzian derivative of these two functions. With $y^2 = x$, we get $\{y,x\} = 3/8x^2$. In the case of interest, $c = 24k$ so

$$T_{xx} = \frac{1}{4x}T_{yy} + \frac{3k}{4x^2}. \tag{4.23}$$

This formula enables us to express T^+ in terms of $T_{yy}|_{y=\sqrt{x}}$ and T^- in terms of $T_{yy}|_{y=-\sqrt{x}}$. The limit in (4.21) becomes

$$\left\langle \left(\frac{1}{4x}T_{yy}|_{y=\sqrt{x}} + \frac{3k}{4x^2}\right)\left(\frac{1}{4x}T_{yy}|_{y=-\sqrt{x}} + \frac{3k}{4x^2}\right)\right\rangle. \tag{4.24}$$

(The idea is that passing to the y-plane eliminates the twist operators $\mathcal{E}$ from the formalism, and replaces $T^\pm(x)$ by a stress tensor evaluated at $y = \pm\sqrt{x}$.) Using the fact that in conformal field theory on the y-plane, $\langle T_{yy}\rangle = 0$ and $\left\langle T_{yy}(y_1)T_{yy}(y_2)\right\rangle = c/2(y_1-y_2)^4$ with $c = 24k$, the correlation function in (4.24) can be evaluated to give θ/x^4 with $\theta = 9k^2/16 + 3k/64$. This gives the last formula in (4.17).

With the aid of the three-point functions in (4.16) and (4.17) and the operator product formula (4.10), one can verify that the candidate four-point function (4.14) has the right singularities precisely if the coefficients A and B are as claimed in (4.15).

Our derivation has in fact been somewhat redundant, since the formulas for three-point functions can largely be deduced from the formula for the $\mathcal{E}\cdot\mathcal{E}$ operator product, and vice-versa.

4.4. The Genus One Partition Function Revisited

We will now explicitly use our formulas for the four-point function of the twist field $\mathcal{E}$ to recover the partition functions of a $k = 1$ or $k = 2$ extremal CFT on a Riemann surface of genus 1.

First we must describe the necessary formalism more precisely. The partition function $Z_C(e_1, e_2, \dots, e_{2g+2})$ of any conformal field theory $\mathcal{W}$ on the hyperelliptic Riemann surface C defined by

$$y^2 = \prod_{i=1}^{2g+2} (x - e_i) \tag{4.25}$$

can indeed be expressed in terms of the genus zero twist field correlation function

$$\big\langle \mathcal{E}(e_1)\mathcal{E}(e_2)\dots\mathcal{E}(e_{2g+2})\big\rangle. \tag{4.26}$$

However, to do so requires an important detail that we have not yet described.

There are at first sight several strange features in relating Z_C to the twist field correlation function:

(1) The dimension of the twist field $\mathcal{E}$ seems to be wrong. The Riemann surface C degenerates as $e_i \to e_j$ for any i and j. As the complex structure of C is invariant under the exchange $e_i \leftrightarrow e_j$, a natural local parameter near the degeneration locus on the moduli space is actually $w = (e_i - e_j)^2$. (For example, this statement can be made precise using eqn. (4.30) below.) One expects Z_C to behave for $e_i \to e_j$ as w^{-k}, since the ground state energy is $-k$. However, as $\mathcal{E}$ has dimension $3k/2$, the behavior of the twist field correlation function as $e_i \to e_j$ is actually $(e_i - e_j)^{-3k} \sim w^{-3k/2}$.

(2) This brings in relief the fact that if k is odd, $\mathcal{E}$ actually has half-integral dimension and so is fermionic. For odd k, the twist field correlation function (4.26) actually changes sign under exchange of e_i and e_j, while Z_C is invariant under this exchange.

(3) The twist field correlation function is most naturally understood as a $3k/2$-differential in each variable. We have written the formulas (4.13) and (4.14) in such a way as to emphasize this. It may not be immediately clear what this means for the partition function Z_C.

The answer to all of these questions has to do with the proper treatment of the conformal anomaly in the holomorphic context. We postpone a full explanation to section

4.5 and for now we merely write down the precise relation between the genus 1 partition function and the twist field correlation function:

$$Z_C(e_1, e_2, \dots, e_4) = 2^{8k}\left(\prod_{1\le i<j\le 4}(e_i - e_j)^k\right)\langle \mathcal{E}(e_1)\mathcal{E}(e_2)\dots\mathcal{E}(e_4)\rangle. \tag{4.27}$$

The factor of 2^{8k} just reflects the way we have normalized $\mathcal{E}$. The factor of $\prod_{i<j}(e_i - e_j)^k$ resolves problems (1) and (2) above. In section 4.5, we will explain the meaning of that factor and its relation to problem (3) (as well as the generalization to higher genus). For now, we merely mention that in eqn. (4.27), the twist field correlation function should be understood as a function, without the factor of $\prod_i (de_i)^{3k/2}$.

Now we will use the general relation (4.27) along with our previous formulas for the twist field correlation functions to analyze the genus 1 partition function for $k = 1, 2$. We make an $SL(2,\mathbb{C})$ transformation to set $e_4 = \infty$ and to fix

$$e_1 + e_2 + e_3 = 0. \tag{4.28}$$

The genus 1 partition function at $k = 1$ is then, using (4.13),

$$Z_C(k=1) = \frac{2^8}{\prod_{i<j}(e_i - e_j)^2}\left((e_1 - e_3)^3(e_2 - e_3)^3 + \text{cyclic}\right) + 24. \tag{4.29}$$

We want to express this in terms of $q = \exp(2\pi i\tau)$, the usual parameter on the genus 1 moduli space, and to show explicitly that Z_C has the expected behavior near $q = 0$. The classical formulas are

$$\begin{aligned} e_1 - e_2 &= \theta_3^4(0,\tau) \\ e_3 - e_2 &= \theta_1^4(0,\tau) \\ e_1 - e_3 &= \theta_2^4(0,\tau), \end{aligned} \tag{4.30}$$

where

$$\begin{aligned} \theta_1(0,\tau) &= \sum_{n\in\mathbb{Z}} q^{(n+1/2)^2/2} \\ \theta_2(0,\tau) &= \sum_{n\in\mathbb{Z}} (-1)^n q^{n^2/2} \\ \theta_3(0,\tau) &= \sum_{n\in\mathbb{Z}} q^{n^2/2}. \end{aligned} \tag{4.31}$$

Expanding (4.29) near $q = 0$, one can readily verify the expected behavior $Z_C(k=1) = q^{-1} + \mathcal{O}(q)$. Since Z_C is manifestly modular-invariant, this implies that Z_C must equal

J. This can also be verified directly by comparing (4.29) to the classical formula for the j-function

$$j(q) = 2^5 \frac{\left((e_1-e_2)^2+(e_2-e_3)^2+(e_3-e_1)^2\right)^3}{\prod_{i<j}(e_i-e_j)^2} \tag{4.32}$$

and recalling that $J = j - 744$.

The case $k = 2$ can be treated similarly. Our previous formula (4.14) for the twist field four-point function implies, via (4.27), a formula for $Z_C(k=2)$. If we take $e_4 \to \infty$ and $e_1 + e_2 + e_3 = 0$, this formula becomes

$$\begin{aligned} Z_C(k=2) =& 2^{16}\left(\frac{(e_1-e_3)^2(e_2-e_3)^2}{(e_1-e_2)^4} + \text{cyclic}\right) \\ &+ 3\cdot 2^{12}\left(\frac{(e_1-e_3)(e_2-e_3)}{(e_1-e_2)^2} + \text{cyclic}\right) + (12k+1). \end{aligned} \tag{4.33}$$

One can expand this near $q = 0$ and verify the expected behavior $Z_C(k=2) = q^{-2} + 1 + \mathcal{O}(q)$. Together with modular invariance, this implies that $Z_C(k=2) = J^2 - 393767$, as one can also verify directly using (4.32).

4.5. Role Of The Conformal Anomaly

Naively speaking, the partition "function" of a conformal field theory on a Riemann surface C depends only on the complex or conformal structure of C.

Actually, because of the conformal anomaly, things are more subtle. The usual way to describe the situation is to endow C with a metric h, and to think of the partition function as a function $Z(h)$ that transforms in a certain prescribed way (depending on the central charge) under conformal rescaling $h \to e^\phi h$ of the metric.

This description is natural in differential geometry. However, it has a few limitations. One is that it really only gives an adequate framework if the left and right central charges c_L and c_R are equal. A second limitation, more crucial for our purposes, is that it is not well-adapted to holomorphic factorization. In a holomorphic CFT, everything will be much simpler if we can incorporate the conformal anomaly in purely holomorphic terms.

An alternative approach is to think of the partition function not as a function but as a section of a suitable line bundle $\mathcal{R}$ over $\mathcal{M}$, the moduli space of complex Riemann surfaces. This approach is developed in detail in [61]. Holomorphic factorization is incorporated naturally in this approach by saying that in a holomorphic CFT, $\mathcal{R}$ is a holomorphic line bundle over $\mathcal{M}$.

More specifically, $\mathcal{R} \cong \mathcal{L}^{c/2}$, where c is the central charge and $\mathcal{L}$ is a fundamental line bundle over $\mathcal{M}$ that is known as the determinant line bundle. $\mathcal{L}$ is defined as follows. Let p be a point in $\mathcal{M}$ corresponding to a Riemann surface C. Then $\mathcal{L}_p$, the fiber of $\mathcal{L}$ at p, is the top exterior power of the vector space $H^0(C, K_C)$ of holomorphic differentials on C. So if C has genus g and $\omega_1, \ldots, \omega_g$ are holomorphic differentials on C, then the expression $\omega_1 \wedge \omega_2 \wedge \ldots \wedge \omega_g$ defines a vector in $\mathcal{L}_p$. In our problem, $c = 24k$, and the partition function is a section of $\mathcal{L}^{12k}$.

On a hyperelliptic Riemann surface

$$y^2 = \prod_{i=1}^{2g+2} (x - e_i), \tag{4.34}$$

we can make this completely explicit. A basis of the space of holomorphic differentials is given by

$$\frac{dx}{y}, \ \frac{x\,dx}{y}, \ \frac{x^2\,dx}{y}, \ldots, \frac{x^{g-1}\,dx}{y}. \tag{4.35}$$

A section of $\mathcal{L}^{12k}$ over the space of hyperelliptic equations is hence an expression of the form

$$\Theta = \left(\frac{dx}{y} \wedge \frac{x\,dx}{y} \wedge \frac{x^2\,dx}{y} \wedge \ldots \wedge \frac{x^{g-1}\,dx}{y} \right)^{12k} F(e_1, \ldots, e_{2g+2}). \tag{4.36}$$

To get a section of $\mathcal{L}^{12k}$ over the moduli space of hyperelliptic Riemann surfaces, Θ must be invariant under the action of $SL(2, \mathbb{C})$ on the space of hyperelliptic equations. This action takes the form

$$\begin{aligned} e_i &\to \frac{ae_i + b}{ce_i + d} \\ x &\to \frac{ax + b}{cx + d} \\ y &\to \frac{y}{(cx + d)^{g+1} \prod_{i=1}^{2g+2} (ce_i + d)^{1/2}}, \end{aligned} \tag{4.37}$$

and the condition for Θ to be invariant is that F transforms by

$$F(e_1, \ldots, e_{2g+2}) \to F(e_1, \ldots, e_{2g+2}) \prod_{i=1}^{2g+2} (ce_i + d)^{-6kg}. \tag{4.38}$$

The twist field correlation function, on the other hand, is of the form

$$\left\langle \mathcal{E}(e_1)\mathcal{E}(e_2)\ldots\mathcal{E}(e_{2g+2}) \right\rangle = \prod_{i=1}^{2g+2} (de_i)^{3k/2} G(e_1, \ldots, e_{2g+2}) \tag{4.39}$$

for some "function" G. $SL(2,\mathbb{C})$ invariance of the correlation function means that G transforms by

$$G(e_1,\dots,e_{2g+2}) \to G(e_1,\dots,e_{2g+2}) \prod_{i=1}^{2g+2} (ce_i+d)^{3k}. \tag{4.40}$$

Comparing (4.38) and (4.40), we see that the two transformation laws are consistent if the relation between F and G is

$$F(e_1,\dots,e_{2g+2}) = A_g\, G(e_1,\dots,e_{2g+2}) \prod_{1\le i<j\le 2g+2} (e_i-e_j)^{3k} \tag{4.41}$$

with some constant A_g. Moreover, this is the most general transformation between F and G that is holomorphic, invariant under permutations of the e_i, and an isomorphism as long as the e_i are all distinct. So it must be correct with some choice of the constant. (The relation between F and G is actually part of a general theory that is briefly indicated below.)

This is not the whole story. The partition function has another important property in genus 1: it can be regarded as an ordinary function, rather than a section of a line bundle. Indeed, the genus 1 partition function can be defined as a trace, $Z(q) = \mathrm{Tr}\, q^{L_0}$, and in this form it manifestly is equal to an ordinary function of q. Yet in the above presentation, the genus 1 partition function is a section of $\mathcal{L}^{12k}$. What reconciles the two points of view is that the line bundle $\mathcal{L}^{12}$ is trivial in genus 1. If we describe a genus 1 Riemann surface C by the hyperelliptic equation $y^2 = \prod_{i=1}^4 (x-e_i)$, then as explained in [61], $\mathcal{L}^{12}$ is trivialized by the section

$$s = \left(\frac{dx}{y}\right)^{12} \prod_{1\le i<j\le 4} (e_i-e_j)^2. \tag{4.42}$$

The point of this formula is that s is holomorphic, $SL(2,\mathbb{C})$-invariant, invariant under permutations of the e_i, and nonzero as long as the e_i are all distinct (so that C is smooth). A power $\mathcal{L}^n$ of the line bundle $\mathcal{L}$ can be trivialized in this fashion precisely if n is an integer multiple of 12 (and this gives one way to understand the fact that holomorphic factorization is possible only if c is an integer multiple of 24). The genus 1 partition function Z understood as an ordinary function is obtained from the section Θ of $\mathcal{L}^{12k}$ by dividing by s^k. Thus

$$Z = \frac{F}{\prod_{1\le i<j\le 4}(e_i-e_j)^{2k}}. \tag{4.43}$$

Combining this with (4.41), we get the relation (4.27) between the genus 1 partition function, understood as a trace, and the correlation function of twist fields:

$$Z = A_1\Big(\prod_{1\le i<j\le 4}(e_i-e_j)\Big)\big\langle \mathcal{E}(e_1)\mathcal{E}(e_2)\mathcal{E}(e_3)\mathcal{E}(e_4)\big\rangle. \tag{4.44}$$

The numerical value of the constant A_1 depends on how the twist field $\mathcal{E}$ has been normalized.

What about $g>1$? What is the best formula depends on how one prefers to treat the conformal anomaly, and there seems to be in the physics literature no standard recipe for doing so in the holomorphic context. To make contact with the genus 2 formulas of Tuite [54], we may proceed as follows. A genus 2 Riemann surface is always a hyperelliptic curve, $y^2=\prod_{i=1}^6(x-e_i)$. In genus 2, we cannot trivialize the line bundle[20] $\mathcal{L}^{12}$, but we can trivialize $\mathcal{L}^{10}$ by an obvious analog of (4.42):

$$\widetilde{s} = \left(\frac{dx}{y}\wedge\frac{x\,dx}{y}\right)^{10}\prod_{1\le i<j\le 6}(e_i-e_j)^2. \tag{4.45}$$

Hence, instead of regarding the genus 2 partition function as a section Θ of $\mathcal{L}^{12k}$, we can regard it as a section $Z=\Theta/\widetilde{s}^k$ of $\mathcal{L}^{2k}$. If we do this, then the relation between Z and the twist field correlation function is the obvious analog of (4.44):

$$Z = A_2\Big(\prod_{1\le i<j\le 6}(e_i-e_j)\Big)\big\langle \mathcal{E}(e_1)\mathcal{E}(e_2)\dots\mathcal{E}(e_6)\big\rangle. \tag{4.46}$$

It can be shown that, in genus 2, a section of $\mathcal{L}^r$ is the same as a Siegel modular function of weight r. So the formula (4.46) is the one we should use if we wish to regard the genus 2 partition function at $c=24k$ as a Siegel modular function of weight $2k$, as was done in [54]. This probably gives the most intuitively appealing formulas in genus 2. For genus greater than 2, we leave the choice of the most useful formalism to the reader.

We conclude by briefly sketching the theoretical context for the formula (4.41). In doing so, we will not limit ourselves to the case of a hyperelliptic Riemann surface, but will consider an arbitrary base curve C_0 and a two-fold cover $C\to C_0$ branched at points $e_1,\dots,e_s$. Let K_i be the cotangent bundle to C_0 at the point e_i. Let $\mathcal{L}_{C_0}$ and $\mathcal{L}_C$ be the determinant lines of C_0 and C respectively. Then as has been shown by P. Deligne, using results in [62], there is a natural isomorphism $\mathcal{L}_C^8\cong\mathcal{L}_{C_0}^{16}\otimes_i K_{C_i}$. If C_0 is of genus zero, then $\mathcal{L}_{C_0}$ is naturally trivial, and the $3k/2$ power of this isomorphism is the map in (4.41). (The isomorphism has a natural square root, relevant if k is odd.)

[20] This fact is consistent with holomorphic factorization at $c=24k$, since the genus 2 partition function cannot be written as a trace and so need not be definable as an ordinary function.

Appendix A. More Extremal Partition Functions

F In section 3.3, we have already described the extremal superconformal partition functions F_{k^*} for $k^* \leq 4$. The purpose of this appendix is to briefly describe these functions, and the corresponding Ramond partition functions H_{k^*}, for $5 \leq k^* \leq 10$.

For $k^* = 5$, we have

$$\begin{aligned} F_5 &= K^5 - 1380K^3 - 10239K^2 + 324871K + 2579929 \\ &= q^{-5/2} + q^{-1} + q^{-1/2} + 1 + 924492q^{1/2} + 185710868q + 11953414737q^{3/2} \\ &\quad + 416513865728q^2 + 9727092442216q^{5/2} + \dots . \end{aligned} \tag{A.1}$$

The corresponding Ramond partition function is

$$H_5 = 143 + 371027968q + 832984743936q^2 + 340658459983872q^3 + 56418463131631616q^4 + \dots . \tag{A.2}$$

The coefficients in these series cannot be expressed as linear combinations of dimensions of monster representations with reasonably small coefficients. But strangely, at least one difference can be so represented; the difference between the coefficient of q in F_5 and half the corresponding coefficient in H_5 is $196884 = 1 + 196883$. (The coefficients in H_5, except the first one, are even because of supersymmetry, so it is natural to divide by 2.) Also, the first coefficient in H_5, namely 143, is the dimension of the smallest nontrivial representation of a sporadic finite group known as the Suzuki group. Unfortunately, the other coefficients are too large, and the representations of the Suzuki group too small, for it to be easy to get convincing evidence of whether the Suzuki group is a symmetry of supergravity at $k^* = 5$.

For $k^* = 6$, the partition functions are

$$\begin{aligned} F_6 &= K^6 - 1656K^4 - 12287K^3 + 618373K^2 + 6487237K - 12026567 \\ &= q^{-3} + q^{-3/2} + q^{-1} + q^{-1/2} + 1 + 3724378q^{1/2} + 1298410586q \\ &\quad + 127852130050q^{3/2} + 6378693040128q^2 + 204402839559265q^{5/2} + \dots \\ H_6 &= 1 + 2589372416q + 12754796707840q^2 \\ &\quad + 9529193701720064q^3 + 2622934801904828416q^4 + \dots . \end{aligned} \tag{A.3}$$

The fact that the leading coefficient in the Ramond sector is 1 suggests that the model may be invariant under a very large discrete symmetry group. The monster $\mathbb{M}$ does not work, since the first non-trivial NS coefficient 3724378 cannot be expressed nicely in terms of dimensions of monster representations. It turns out that another large sporadic group,

the baby monster $\mathbb{B}$, is a better candidate. The dimensions $r_1, \ldots, r_{12}$ of the first 12 irreducible representations of $\mathbb{B}$ are listed in Table 3. In terms of those dimensions, we find that the first two non-trivial NS coefficients in the partition functions can be expressed as follows

$$\begin{aligned} 3724378 &= 7r_1 + 4r_2 + 3r_3 + 3r_4 \\ 1298410586 &= 14r_1 + 16r_2 + 7r_3 + 8r_4 + 4r_6 + 3r_7 + r_8 + 2r_9. \end{aligned} \tag{A.4}$$

These formulas are not unique, as the r_i obey linear relations with small coefficients, for example $r_1 + r_3 + r_5 + r_8 = r_2 + r_4 + r_9$. If we let h_1 denote the first non-trivial coefficient in H_6, then it turns out that $h_1/2 = f_1 - f_{1/2} = 1298410586 - 3724378$. So by subtracting the formulas in (A.4), $h_1/2$ can also be expressed as a positive linear combination of the r_i with fairly small coefficients. If is also true that $h_2/2 = f_2 - f_1 + f_{1/2}$, and so is consistent with $\mathbb{B}$ symmetry if f_2 can be suitably expressed in terms of the r_i.

These results are suggestive of baby monster symmetry at $k^* = 6$, though they are perhaps a little less striking than the evidence presented in section 3.3 for monster symmetry at $k^* = 4$, since the required coefficients are larger and the dimensions involved are smaller. Also, it is harder to continue this analysis beyond the first few coefficients, as many representations of $\mathbb{B}$ come in.

r_1	1
r_2	4371
r_3	96255
r_4	1139374
r_5	9458750
r_6	9550635
r_7	63532485
r_8	347643114
r_9	356054375
r_{10}	1407126890
r_{11}	3214743741
r_{12}	4221380670

Table 3. Presented here from [53] are the dimensions r_i of the ith irreducible representation of the baby monster group $\mathbb{B}$, for $i = 1, \ldots, 12$.

We present the remaining cases with little comment. For $k^* = 7$, we have

$$\begin{aligned} F_7 &= K^7 - 1932K^5 - 14335K^4 + 988051K^3 + 11525041K2 - 75824563K - 840705550 \\ &= q^{-7/2} + q^{-2} + q^{-3/2} + q^{-1} + q^{-1/2} + 2 + 13404883q^{1/2} \\ &\quad + 7740446996q + 1126452195714q^{3/2} + 78170641348884q^2 + \dots . \\ H_7 &= 262 + 15394525184q + 156260255891456q^2 + 203203950584774656q^3 + \dots . \end{aligned} \tag{A.5}$$

For $k^* = 8$, the analogous formulas are

$$\begin{aligned} F_8 &= K^8 - 2208K^6 - 16383K^5 + 1433905K^4 + 17693341K^3 - 213343055K^2 \\ &\quad - 3164679732K - 2780845557 \\ &= q^{-4} + q^{-5/2} + q^{-2} + q^{-3/2} + q^{-1} + 2q^{-1/2} + 3 + 44146598q^{1/2} \\ &\quad + 40700662036q + 8516908978515q^{3/2} + \dots \\ H_8 &= -213 + 80651894784q + 1605169778655232q^2 + 3496922597386551296q^3 + \dots . \end{aligned} \tag{A.6}$$

This is the first case in which the leading coefficient in H is negative, so that it is necessary to add an integer to F and H.

For $k^* = 9$, we find

$$\begin{aligned} F_9 &= K^9 - 2484K^7 - 18431K^6 + 1955935K^5 + 24992137K^4 - 445606619K^3 \\ &\quad - 7443672266K^2 + 1774761946K + 223898812203 \\ &= q^{-9/2} + q^{-3} + q^{-5/2} + q^{-2} + q^{-3/2} + 2q^{-1} + 3q^{-1/2} + 3 + 135149374q^{1/2} \\ &\quad + 193216791918q + 56847816152503q^{3/2} + \dots \\ H_9 &= 453 + 381161021440q + 14282018665201664q^2 + 5051928839165624320 0q^3 + \dots \end{aligned} \tag{A.7}$$

Finally, for $k^* = 10$, we find

$$\begin{aligned} F_{10} &= K^{10} - 2760K^8 - 20479K^7 + 2554141K^6 + 33421429K^5 - 793639831K^4 \\ &\quad - 14145708496K^3 + 30831695165K^2 + 1166011724825K + 2482063616019 \\ &= q^{-5} + q^{-7/2} + q^{-3} + q^{-5/2} + q^{-2} + 2q^{-3/2} + 3q^{-1} + 3q^{-1/2} + 3 \\ &\quad + 389274233q^{1/2} + 842231630010q + 341925580784341q^{3/2} + \dots \\ H_{10} &= -261 + 1652836102144q + 112692628289650688q^2 + 6305205669016140021 76q^3 + \dots . \end{aligned} \tag{A.8}$$

This gives a second example in which the leading coefficient of H is negative, so that it is necessary to add an integer.

References

[1] H. Leutwyler, "A 2+1 Dimensional Model For The Quantum Theory Of Gravity," Nuovo Cim. **42A** (1966) 159.

[2] E. J. Martinec, "Soluble Systems In Quantum Gravity," Phys. Rev. **D30** (1984) 1198.

[3] S. Deser, R. Jackiw, and G. 't Hooft, "Three-Dimensional Einstein Gravity: Dynamics of Flat Space," Annals Phys. **152** (1984) 220.

[4] S. Carlip, "Conformal Field Theory, $(2+1)$-Dimensional Gravity, and the BTZ Black Hole," gr-qc/0503022.

[5] A. Achúcarro and P. Townsend, "A Chern-Simons Action for Three-Dimensional anti-De Sitter Supergravity Theories," Phys. Lett. **B180** (1986) 89.

[6] E. Witten, "(2+1)-Dimensional Gravity as an Exactly Soluble System," Nucl. Phys. **B311** (1988) 46.

[7] A. Ashtekar, "New Variables For Classical And Quantum Gravity," Phys. Rev. Lett. **57** (1986) 2244-2247.

[8] J. Teschner, "On The Relation Between Quantum Liouville Theory And The Quantum Teichmuller Spaces," Int. J. Mod. Phys. **A19S2** (2004) 459-477, hep-th/0303149.

[9] J. Teschner, "From Liouville Theory To The Quantum Gravity Of Riemann Surfaces" (International Congress on Mathematical Physics 2003), hep-th/0308031.

[10] J. Teschner, "A Lecture On The Liouville Vertex Operators," Int. J. Mod. Phys. **A19S2** (2004) 436-458, hep-th/0303150.

[11] J. Teschner, "An Analog Of A Modular Functor From Quantized Teichmuller Theory," math/0510174.

[12] M. Bañados, C. Teitelboim, and J. Zanelli, "The Black Hole In Three-Dimensional Spacetime," Phys. Rev. Lett. **69** (1992) 1849-1851, hep-th/9204099.

[13] A. Strominger, "Black Hole Entropy From Near Horizon Microstates," JHEP 9802:009 (1998), hep-th/9712251.

[14] D. Birmingham, I. Sachs, and S. Sen, "Entropy Of Three-Dimensional Black Holes In String Theory," Phys. Lett. **B424** (1998) 275-280, hep-th/9801019.

[15] J. Maldacena, "The Large N Limit Of Superconformal Field Theories And Supergravity," Adv. Theor. Math. Phys. **2** (1998) 231-252, hep-th/9711200.

[16] C. G. Callan, Jr., S. B. Giddings, J. A. Harvey, and A. Strominger, "Evanescent Black Holes," Phys. Rev. **D45** (1992) 1005-1009, hep-th/9111056.

[17] N. Goheer, M. Kleban, and L. Susskind, "The Trouble With de Sitter Space," JHEP 0307:066 (2003), hep-th/0212209.

[18] E. Witten, "Quantum Gravity In de Sitter Space," hep-th/0106109.

[19] A. Strominger, "The dS/CFT Correspondence," JHEP 0110:034 (2001), hep-th/0106113.

[20] S. Kachru, R. Kallosh, A. Linde, and S. Trivedi, "de Sitter Vacua In String Theory," Phys. Rev. D68:046005,2003, hep-th/0301240.

[21] J. D. Brown and M. Henneaux, "Central Charges In The Canonical Realization Of Asymptotical Symmetries: An Example From Three-Dimensional Gravity," Commun. Math. Phys. **104** (1986) 207-226.

[22] A. B. Zamolodchikov, "Irreversibility of the Flux of the Renormalization Group in a 2D Field Theory," JETP Lett. **43** (1986) 430.

[23] A. N. Schellekens, "Meromorphic $c = 24$ Conformal Field Theories," Commun. Math. Phys. **153** (1993)159-186, hep-th/9205072.

[24] I. B. Frenkel, J. Lepowsky, and A. Meurman, "A Natural Representation of the Fischer-Griess Monster With the Modular Function J As Character," Proc. Natl. Acad. Sci. USA **81** (1984) 3256-3260.

[25] I. B. Frenkel, J. Lepowsky, and A. Meurman, *Vertex Operator Algebras And The Monster* (Academic Press, Boston, 1988).

[26] L. Dixon, P. Ginsparg, and J. Harvey, "Beauty And The Beast: Superconformal Symmetry In A Monster Module," Commun. Math. Phys. **119** (1988) 221-241.

[27] L. Dolan, P. Goddard, and P. Montague, "Conformal Field Theory, Triality, and the Monster Group," Phys. Lett. **B236** (1990) 165-172.

[28] R. Borcherds, "Automorphic Forms And Lie Algebras," in *Current Developments In Mathematics* (International Press, Boston, 1997).

[29] J. McKay, "The Essentials of Monstrous Moonshine," *Groups and Combinatorics - in Memory of M. Suzuki*, Adv. Studies in Pure Math. **32** (2001) 347-343.

[30] T. Gannon, "Moonshine Beyond The Monster: The Bridge Connecting Algebra, Modular Forms, and Physics" (Cambridge University Press, 2006).

[31] G. Höhn, "Selbstduale Vertexoperatorsuperalgebren und das Babymonster," Ph.D. thesis (Bonn 1995), Bonner Mathematische Schriften 286 (1996), 1-85, arXiv:0706.0236.

[32] G. Höhn, "Conformal Designs based on Vertex Operator Algebras," arXiv:math/0701626.

[33] M. Jankiewicz and T. Kephart, "Modular Invariants And Fischer-Griess Monster," arXiv:math-ph/0608001.

[34] R. Dijkgraaf, J. Maldacena, G. Moore, and E. Verlinde, "A Black Hole Farey Tail," hep-th/0005003.

[35] J. F. Duncan, "Super-Moonshine For Conway's Largest Sporadic Group," arXiv:math/0502267.

[36] S. Deser, R. Jackiw, and S. Templeton, "Topologically Massive Gauge Theories," Annals Phys. **140** (1982) 372-411.

[37] G. Moore and N. Seiberg, "Taming The Conformal Zoo," Phys. Lett. **B220** (1989) 422.

[38] E. Witten,"$SL(2,\mathbb{Z})$ Action On Three-Dimensional Conformal Field Theories With Abelian Symmetry," hep-th/0307041.

[39] S. Carlip "Quantum Gravity In 2+1 Dimensions: The Case Of A Closed Universe," Living Rev. Rel. **8** (2005) 1, gr-qc/0409039.

[40] J. Hartle and S. B. Hawking, "Wavefunction Of The Universe," Phys. Rev. **D28** (1983) 2960.

[41] D. Friedan and S. Shenker, "The Analytic Geometry Of Two-Dimensional Conformal Field Theory," Nucl. Phys. **B281** (1987) 509.

[42] S. Gukov, E. Martinec, G. W. Moore, and A. Strominger, "An Index for 2-D Field Theories with Large $\mathcal{N} = 4$ Superconformal Symmetry," hep-th/0404023.

[43] J. Thompson, "Some Numerology Between the Fischer-Griess Monster and the Elliptic Modular Function," Bull. London Math. Soc. 11 (1979), 352-353.

[44] J. H. Conway and S. P. Norton, "Monstrous Moonshine," Bull. London Math. Soc. **11** (1979) 308-339.

[45] N. Koblitz, *Introduction To Elliptic Curves And Modular Forms* (Springer-Verlag, 1984).

[46] A. Maloney and E. Witten, to appear.

[47] D. Birmingham and S. Sen, "Exact Black Hole Entropy Bound In Conformal Field Theory," hep-th/0008051.

[48] L. Dixon, J. A. Harvey, C. Vafa, and E. Witten, "Strings On Orbifolds," Nucl. Phys. **B261** (1985) 678-686.

[49] M. P. Tuite, "On The Relation Between Generalized Moonshine And The Uniqueness Of The Moonshine Module," Commun. Math. Phys. **166** (1995) 495-532, hep-th/9305057.

[50] J. Conway and N. A. Sloane, *Sphere Packings, Lattices, and Groups* (Grundlehren der Mathematischen Wissenschaften, 1998).

[51] G. Nebe, "Some Cyclo-Quaternionic Lattices," J. Algebra **199** (1998) 472-298.

[52] F. Gliozzi, D. I. Olive, and J. Scherk, "Supersymmetry, Supergravity Theories, And The Dual Spinor Model," Nucl. Phys. **B122** (1977) 253-290.

[53] J. H. Conway, R. T. Curtis, S. P. Norton, R. A. Parker, and R. A. Wilson, *Atlas Of Finite Groups* (Oxford University Press, 1985).

[54] M. P. Tuite, "Genus Two Meromorphic Conformal Field Theory," math.qa/9910136.

[55] G. Mason and M. P. Tuite, "On Genus Two Riemann Surfaces Formed From Sewn Tori," Commun. Math. Phys. **270** (2007) 587-634, math.qa/0603088.

[56] Al. B. Zamolodchikov, "Correlation Functions Of Spin Operators In The Ashkin-Teller Model And The Scalar Field On A Hyperelliptic Surface," Soviet Phys. JETP **63** (1986) 1061-1066.

[57] V. G. Knizhnik, "Analytic Fields On Riemann Surfaces," Commun. Math. Phys. **112** (1987) 567.

[58] L. Dixon, D. Friedan, E. Martinec, and S. Shenker, "The Conformal Field Theory Of Orbifolds," Nucl. Phys. **B282** (1987) 13-73.

[59] S. Hamidi and C. Vafa, "Interactions On Orbifolds," Nucl. Phys. **B279** (1987) 465.

[60] A. B. Zamolodchikov, "Infinite Additional Symmetries In Two-Dimensional Conformal Quantum Field Theory," Theor. Math. Phys. **65** (1985) 1205-1213.

[61] G. Segal, "The Definition Of Conformal Field Theory," in U. Tillmann, ed. *Topology Geometry, and Quantum Field Theory* (Cambridge University Press, 2004).

[62] P. Deligne, "Le Déterminant de La Cohomologie," in *Current Trends In Arithmetical Algebraic Geometry*, Contemp. Math. **67** (Amer. Math. Soc., Providence, 1987).

arXiv:yymm.nnnn

Quantum Gravity Partition Functions In Three Dimensions

arXiv:0712.0155v1 [hep-th] 2 Dec 2007

Alexander Maloney[1] and Edward Witten[2]

[1]McGill Physics Department, 3600 rue University, Montréal, QC H3A 2T8, Canada
[2]School of Natural Sciences, Institute for Advanced Study, Einstein Dr., Princeton, NJ 08540

Abstract

We consider pure three-dimensional quantum gravity with a negative cosmological constant. The sum of known contributions to the partition function from classical geometries can be computed exactly, including quantum corrections. However, the result is not physically sensible, and if the model does exist, there are some additional contributions. One possibility is that the theory may have long strings and a continuous spectrum. Another possibility is that complex geometries need to be included, possibly leading to a holomorphically factorized partition function. We analyze the subleading corrections to the Bekenstein-Hawking entropy and show that these can be correctly reproduced in such a holomorphically factorized theory. We also consider the Hawking-Page phase transition between a thermal gas and a black hole and show that it is a phase transition of Lee-Yang type, associated with a condensation of zeros in the complex temperature plane. Finally, we analyze pure three-dimensional supergravity, with similar results.

email: maloney@physics.mcgill.ca, witten@ias.edu

Contents

1. Introduction . . . 1
2. Known Contributions To The Path Integral . . . 5
   2.1. Classification Of Solutions . . . 6
   2.2. Evaluation Of The Partition Function . . . 13
   2.3. The General Form Of The Partition Function . . . 21
3. Computing the Sum over Geometries . . . 22
   3.1. The Regularized Sum . . . 23
   3.2. Poisson Resummation . . . 26
   3.3. The $\hat{n} = 0$ Mode . . . 29
   3.4. $\hat{n} \neq 0$ Modes . . . 33
   3.5. Aside: the Tree Level Sum over Geometries . . . 35
4. Possible Interpretations . . . 36
   4.1. Cosmic Strings . . . 36
   4.2. Doubled Sum Over Geometries . . . 38
5. Black Hole Entropy and its Corrections . . . 42
   5.1. Subleading Corrections to the Bekenstein-Hawking Formula . . . 43
   5.2. The Rademacher Expansion for Holomorphic Partition Functions . . . 46
6. The Hawking-Page Phase Transition . . . 48
   6.1. Review: Properties of J . . . 52
   6.2. Properties of $T_n J$ . . . 54
   6.3. Zeroes of Modular Invariant Partition Functions . . . 56
7. Supergravity Partition Functions . . . 58
   7.1. Formalism . . . 58
   7.2. The Computation . . . 62
   7.3. Phase Transitions . . . 67

1. Introduction

This paper is devoted to describing some explicit computations relevant to three-dimensional pure quantum gravity with negative cosmological constant. The classical action can be written

$$I = \frac{1}{16\pi G} \int d^3x \sqrt{g} \left(R + \frac{2}{\ell^2} \right). \tag{1.1}$$

Three-dimensional gravity with a cosmological constant was first studied in [1]. Some early milestones in the study of this theory, both of them special to the case of negative cosmological constant, have been the construction of an asymptotic Virasoro symmetry [2] and the recognition that the theory admits black holes [3,4]. The asymptotic Virasoro

symmetry is now understood as part of the structure of a dual two-dimensional conformal field theory [5] and this structure together with modular invariance can be used to determine the entropy of a black hole of asymptotically large mass [6]. For a recent review of the extensive work on this subject, with references, see [7].

None of the results mentioned in the last paragraph are special to the case of *pure* three-dimensional gravity, without additional fields. Our intent here, however, is to study this minimal theory, with the goal of computing its exact energy levels in a spacetime that is asymptotic at spatial infinity to the classical "vacuum" of three-dimensional Anti de Sitter space, AdS_3. As always in General Relativity, a well-posed problem is obtained by specifying what the world should look like at spatial infinity – in this case we ask that it should coincide with AdS_3 at infinity – and then analyzing all possible "interiors." Our problem in this paper is to compute the precise quantum energy levels that arise. We perform a computation based on known concepts, and in a sense that we will explain, this computation is not successful. The reasons for this failure are not clear and we consider several hypotheses.

It is convenient to summarize the spectrum of energy levels in the form of a trace $\mathrm{Tr}\,\exp(-\beta H)$, where H is the Hamiltonian and β is a positive real number (or more generally a complex number with positive real part). As usual in General Relativity, the Hamiltonian is defined via the ADM procedure in terms of the leading behavior at spatial infinity of the correction to the pure AdS_3 metric. It was indeed by carefully examining this procedure that the asymptotic Virasoro symmetry (which generalizes the obvious conservation laws such as conservation of energy) was discovered [2].

There is also a conserved angular momentum J which generates a rotation at infinity of the asymptotically AdS_3 spacetime and commutes with H. Consequently, one can introduce an additional parameter θ and try to compute a more general partition function:

$$Z(\beta,\theta) = \mathrm{Tr}\,\exp(-\beta H - i\theta J). \tag{1.2}$$

This partition function is naturally computed via a Euclidean path integral. According to the standard recipe, the integral is taken over Euclidean three-geometries that are conformal at infinity to a two-torus Σ with modular parameter $\tau = \theta/2\pi + i\beta$. We write $|dz|^2$ for a flat metric on this torus, where z is a complex coordinate subject to the identifications $z \to z+1$, $z \to z+\tau$, and introduce another coordinate $u > 0$, such that conformal infinity will be the region $u \to 0$. Then the metric should behave for $u \to 0$ like

$$ds^2 = \frac{|dz|^2 + du^2}{u^2} \tag{1.3}$$

plus subleading terms.

The Euclidean path integral also has a natural interpretation in the dual two-dimensional CFT [5]: it is the genus one partition function of this dual theory, on the two-dimensional surface Σ. In fact, by definition, the Hamiltonian and momentum H and P of the dual CFT coincide with H and J, the Hamiltonian and angular momentum of the original theory in AdS_3. This dual interpretation is informative, but is not really needed to motivate the computation that we perform.

This Euclidean path integral is a formal recipe, for various reasons. One problem is that, in general, the Euclidean quantum gravity path integral is not convergent because the action is not bounded below [8]. The only known way to deal with this problem is to expand around a classical solution; in doing so, one can obtain a meaningful result at least in perturbation theory. There is no clearly established claim in the literature that topologies that do not admit classical solutions do not contribute to the Euclidean path integral. But there is also no known method to evaluate their contributions.

It turns out that in the present case, one can describe the classical solutions completely. This is explained in Sec. 2.1. The classical solutions are precisely the ones considered in [9] and in [10] (in studying the elliptic genus of certain AdS_3 models derived from string theory). Moreover, perturbation theory around a classical solution terminates with the one-loop term, and that term can be easily evaluated by adapting the arguments of [2]. This is the content of Sec. 2.2.

We are therefore in a position to write down the complete sum of *known* contributions to the path integral. This is done in Sec. 3. The sum turns out to require some regularization, and we use what we believe is a natural regularization, analogous to zeta function regularization [11]. We ultimately obtain an explicit, though complicated, formula for the sum of known contributions to $Z(\beta,\theta)$. However – and this is our main result – the sum is not physically sensible: it cannot be written as $\mathrm{Tr}\,\exp(-\beta H - i\theta J)$ for any commuting operators H, J in a Hilbert space. (According to [12], this possibility was also conjectured by S. Minwalla.)

We do not know the correct interpretation of this result. In Sec. 4, we discuss some possibilities. One possibility is that the *minimal* theory that we postulate here actually does not exist. There are many quantum theories that do exist that look semiclassically like three-dimensional gravity coupled to additional fields; indeed, there are a plethora of

known string-derived models, such as the ones studied in [10]. However, it is not clear[1] that there exists nonperturbatively a minimal theory that one would want to call pure gravity. One possible interpretation of our result is to indicate that this minimal theory does not exist.

The other possibility, broadly speaking, is that there are additional contributions to the path integral beyond the known ones that we evaluate in Secs. 2 and 3. To be concrete, we consider in Sec. 4 two scenarios. One is that the minimal theory of three-dimensional gravity, in addition to the known BTZ black holes and Brown-Henneaux boundary excitations, also describes cosmic strings. The motivation for this proposal is that known models of three-dimensional quantum gravity, such as the string-based models considered in [10], do always have cosmic strings. Perhaps this is also true for the "minimal" theory, if it exists.

The second scenario is that in addition to the real saddle points that are classified and evaluated in Secs. 2 and 3, an exact description of the theory should also include complex saddle points. We describe a specific scenario in which the inclusion of complex saddle points leads to the holomorphically factorized partition function that was proposed (based on highly speculative reasoning) in [13]. This involves a doubled sum over saddle points similar to what is assumed in [12,14]. The resulting partition function is consistent with an interpretation as Tr $\exp(-\beta H - i\theta J)$ with some Hilbert space operators H, J.

Though we are not able to put this proposal in a convincing form, we do uncover one interesting fact, which concerns the semiclassical behavior of the partition function (assumed to be holomorphically factorized and extremal) as the gravitational coupling G goes to zero with fixed AdS radius ℓ. In this semiclassical limit, the partition function $Z = \mathrm{Tr}\, \exp(-\beta H - i\theta J)$ is always dominated, as long as β and θ are real, by a real saddle point corresponding to a real geometry. The partition function can be dominated by a classical saddle point, but only if one asks a more exotic question with complex values of β and θ.

In Secs. 5 and 6 we discuss questions of black hole physics in three dimensions. In section 5 we discuss black hole entropy: a theory of pure quantum gravity, assuming it

[1] The problem may be analogous to trying to define a minimal string theory in four dimensions. There are many theories that macroscopically are four-dimensional string theories, such as theories obtained by compactification to four dimensions from the critical dimension, or gauge theories with flux tubes or vortex lines, but it is plausible that there is nothing that should sensibly be called a minimal four-dimensional quantum string theory.

exists, should give a proper microscopic accounting of the Bekenstein-Hawking entropy of the BTZ black hole. In the semi-classical limit this entropy is just the horizon area. The computations of section 2, however, allow us to determine the perturbative corrections to this semi-classical result, which are described in section 5.1. In section 5.2 we show that an infinite series of corrections can be reproduced in a holomorphically factorized theory.

In investigating these holomorphically factorized theories, we noticed an additional interesting phenomenon that is the subject of Sec. 6. Three-dimensional quantum gravity, as a function of β, exhibits in the semi-classical limit the Hawking-Page phase transition [15] between a thermal gas (in this case, a gas of Brown-Henneaux boundary excitations) and a black hole. At first sight, it is puzzling how such a phase transition can be compatible with holomorphic factorization. We show, however, that this question has a natural answer, in terms of a condensation on the phase boundary of Lee-Yang zeroes of the partition function.

Finally, in Sec. 7, we extend our analysis to pure $\mathcal{N} = 1$ supergravity, obtaining results similar to what we find in the bosonic case.

2. Known Contributions To The Path Integral

Our task in this section is two-fold. The first step is to classify Euclidean solutions of Einstein's equations, with negative cosmological constant, with the asymptotic behavior described in the introduction. We write M for the three-dimensional spacetime, and Σ for its conformal boundary; as in the introduction, Σ is a Riemann surface of genus 1. We assume that M is smooth, that the metric on M is complete, and that Σ is the only "end" of M. According to the standard logic of Euclidean quantum gravity, known contributions to the path integral have these properties; it may be appropriate physically to relax them, but we do not know how.

We proceed in two steps. In Sec. 2.1, we describe the possible choices of M. The result is actually well-known in the theory of hyperbolic three-manifolds. In Sec. 2.2, we evaluate the contribution to the partition function of a particular M. The sum over different choices of M is postponed to Sec. 3. In Sec. 2.3, we describe the general form of the partition function in a theory with finite entropy.

2.1. Classification Of Solutions

The automorphism group of AdS_3 is $SO(3,1)$, which is the same as $SL(2,\mathbb{C})/\mathbb{Z}_2$. We may write the metric on a dense open subset of AdS_3 as[2]

$$ds^2 = \frac{|dz|^2 + du^2}{u^2}, \qquad u > 0, \qquad z \in \mathbb{C}. \tag{2.1}$$

If we combine the (z,u) coordinates into a single quaternion $y = z + ju$, the action of an element $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in SL(2,\mathbb{C})$ can be written succinctly as

$$y \to (ay+b)\,(cy+d)^{-1}\,. \tag{2.2}$$

In this expression the element $\left(\begin{smallmatrix} -1 & 0 \\ 0 & -1 \end{smallmatrix}\right) \in SL(2,\mathbb{C})$ acts trivially, so (2.2) actually describes the action of $SL(2,\mathbb{C})/\mathbb{Z}_2$ on AdS_3. In general, any classical solution M of three-dimensional gravity with negative cosmological constant takes the form AdS_3'/Γ, where Γ is a discrete subgroup of $SO(3,1)$ and AdS_3' is the part of AdS_3 on which Γ acts discretely.

The conformal boundary of M can be constructed as follows. First of all, the conformal boundary of AdS_3 is a two-sphere, which one can think of as $\mathbb{CP}^1$, acted on by $SL(2,\mathbb{C})$ in the usual way. This $\mathbb{CP}^1$ may be regarded as the complex z-plane in (2.1) at $u \to 0$ plus a point at infinity. From (2.2), one can see that $SL(2,\mathbb{C})$ acts on this $\mathbb{CP}^1$ in the familiar fashion

$$z \to \frac{az+b}{cz+d}. \tag{2.3}$$

To construct the conformal boundary of $M = \mathrm{AdS}_3'/\Gamma$, one first throws away a certain subset of $\mathbb{CP}^1$ in a neighborhood of which the discrete group Γ acts badly. This set is closed, so its complement is an open subset $U \subset \mathbb{CP}^1$.

The discrete group Γ acts freely on U, and the conformal boundary of M is the quotient

$$\Sigma = U/\Gamma. \tag{2.4}$$

Since the action of $SL(2,\mathbb{C})$ preserves the holomorphic structure of $\mathbb{CP}^1$, U/Γ carries a natural holomorphic structure and the isomorphism (2.4) is valid holomorphically, not just topologically.

Given this, let us investigate the condition for Σ to be of genus 1. Σ is topologically a two-torus, so its fundamental group is $\pi_1(\Sigma) = \mathbb{Z} \oplus \mathbb{Z}$. It follows from eqn. (2.4) that the

[2] We set $\ell = 1$ unless otherwise indicated.

fundamental group of U is a subgroup of the fundamental group of Σ. Possible subgroups $\pi_1(U) \subset \mathbb{Z} \times \mathbb{Z}$ are of three types:

(i) $\pi_1(U)$ may be a subgroup of $\pi_1(\Sigma)$ of finite index, isomorphic to $\mathbb{Z} \oplus \mathbb{Z}$. (A special case is $\pi_1(U) = \pi_1(\Sigma)$.)

(ii) $\pi_1(U)$ may consist of multiples of a single non-zero vector $x \in \pi_1(\Sigma)$, in which case $\pi_1(U) \cong \mathbb{Z}$.

(iii) $\pi_1(U)$ may be trivial.

In case *(i)*, U is a finite cover of Σ, and therefore is itself a Riemann surface of genus 1. However, a Riemann surface of genus 1 is not isomorphic to an open subset of $\mathbb{CP}^1$. So case *(i)* cannot arise.

Cusp Geometry

In case *(iii)*, U is the universal cover of Σ, and so is isomorphic to $\mathbb{C}$ (or $\mathbb{R}^2$). The holomorphic structure of $\mathbb{C}$ is unique up to isomorphism. $\mathbb{C}$ is isomorphic holomorphically to an open subset of $\mathbb{CP}^1$ in essentially only one way: it is the complement of one point, say the point at $z = \infty$. The subgroup of $SL(2, \mathbb{C})$ that leaves fixed the point at infinity consists of the triangular matrices

$$\begin{pmatrix} \lambda & w \\ 0 & \lambda^{-1} \end{pmatrix}. \tag{2.5}$$

The point $z \in \mathbb{C}$ corresponds to $\begin{pmatrix} z \\ 1 \end{pmatrix} \in \mathbb{CP}^1$, and a triangular matrix acts by $z \to \lambda^2 z + \lambda w$.

Since U is simply-connected, Γ must be isomorphic to $\mathbb{Z} \oplus \mathbb{Z}$ (in order to get the right fundamental group for Σ). Any discrete group of triangular matrices that is isomorphic to $\mathbb{Z} \oplus \mathbb{Z}$ is generated by two strictly triangular matrices

$$\begin{pmatrix} 1 & a \\ 0 & 1 \end{pmatrix}, \quad \begin{pmatrix} 1 & b \\ 0 & 1 \end{pmatrix}, \tag{2.6}$$

where a and b are complex numbers that are linearly independent over $\mathbb{R}$. (If a and b are linearly dependent over $\mathbb{R}$, then the group generated by these matrices is not discrete, since a suitable linear combination $ma + nb$, for $m, n \in \mathbb{Z}$, can be arbitrarily small.) Up to conjugacy by a diagonal matrix, the only invariant of such a group is the ratio b/a. Therefore, we can reduce to the case $a = 1$, $b = \tau$, and moreover by taking $b \to -b$ (which does not affect the group generated by the two matrices) we can assume that $\operatorname{Im} \tau > 0$.

We have therefore arrived precisely at the group of symmetries

$$z \to z + m + n\tau, \ \ m, n \in \mathbb{Z} \tag{2.7}$$

of the complex z-plane. The quotient is a genus 1 surface Σ with an arbitrary τ-parameter. The three-manifold $M = \mathrm{AdS}_3/\Gamma$ can be described very concretely and is a standard example. The (u, z) coordinates in (2.1) cover precisely the subspace AdS_3' on which Γ acts "nicely." So $M = \mathrm{AdS}_3'/\Gamma$ is given by the metric (2.1), now subject to the identification (2.7).

But M does not obey the conditions described at the beginning of this section. M is smooth and has a complete Einstein metric. However, in addition to the "end" at $u = 0$, which is the one required by our boundary condition, M also has a second "end" at $u = \infty$. Our problem was to classify Einstein manifolds with only one end; M does not qualify.

The second end of M is one at which the metric of Σ collapses to zero (rather than blowing up, as it does for $u \to 0$). An end of this kind is known as a "cusp." In the theory of hyperbolic three-manifolds, one often considers hyperbolic metrics with such cusps. One of the main results about them is that, under suitable conditions, such a metric can be slightly perturbed so as to eliminate the cusp. In the present context, this gives the sort of metric that we will describe momentarily.

Physically, we cannot be certain that omitting the case of the "cusp" is the correct thing to do. However, restricting to cuspless metrics appears to be the right thing to do in known examples of AdS_3 theories. And pragmatically, we believe that from what is known of three-dimensional gravity, it would be difficult to give a sensible procedure for evaluating the contribution of the spacetime with the cusp. The reason for the last statement is that non-trivial one-cycles (loops in Σ) become sub-Planckian in length near a cusp, so a semiclassical treatment is not valid.

Semiclassical Geometries

The last case to consider – case *(ii)* – is that the fundamental group of U is $\mathbb{Z}$. This means that topologically U is $\mathbb{R} \times S^1$ (S^1 is a circle). The holomorphic structure of U is then uniquely determined to be that of the z-plane minus a point, which we may as well take to be the point at $z = 0$. The subgroup of triangular matrices that preserve the point $z = 0$ is simply the group of diagonal matrices.

Therefore Γ is a discrete subgroup of the group of diagonal matrices. It cannot be a finite group (or the first Betti number of $\Sigma = U/\Gamma$ would be 1, while the desired value is 2). This being so, there are essentially two cases to consider.

First, Γ may be isomorphic to $\mathbb{Z}$, generated by a matrix of the form

$$W = \begin{pmatrix} q & 0 \\ 0 & q^{-1} \end{pmatrix} \in SL(2,\mathbb{C}). \tag{2.8}$$

By exchanging the two eigenvalues we can assume that $|q| < 1$. (If $|q| = 1$, then either q is a root of unity and the group generated by W is a finite group, or q is not a root of unity and the subgroup of $SL(2,\mathbb{C})$ generated by W is not discrete.) Alternatively, Γ may be isomorphic to $\mathbb{Z} \times \mathbb{Z}_n$, generated by W together with

$$Y = \begin{pmatrix} \exp(2\pi i/n) & 0 \\ 0 & \exp(-2\pi i/n) \end{pmatrix}, \tag{2.9}$$

with some integer n.

Let us first consider the case that $\Gamma = \mathbb{Z}$. Then $\Sigma = U/\Gamma$ is obtained from the complex z-plane by throwing away the point $z = 0$ and dividing by the group generated by W. It is convenient to write $z = \exp(2\pi i w)$, so that w is defined modulo

$$w \to w + 1 \tag{2.10}$$

and W acts by

$$w \to w + \frac{\log q}{2\pi i}. \tag{2.11}$$

The quotient of the w-plane by (2.10) and (2.11) is a Riemann surface of genus 1, as required. The complex modulus of this surface is $\tau = \frac{\log q}{2\pi i}$, i.e. it is given by $q = e^{2\pi i \tau}$. More generally, however, the modulus of this Riemann surface is defined only up to $\tau \to (a\tau + b)/(c\tau + d)$ with integers a, b, c, d obeying $ad - bc = 1$. Therefore, we will get an equivalent Riemann surface if

$$q = \exp(2\pi i (a\tau + b)/(c\tau + d)) \tag{2.12}$$

for such a, b, c, d.

One might conclude, therefore, that we get a three-manifold obeying the required conditions for every choice of a, b, c, d. This is not quite the case, for two reasons. First, an overall sign change of a, b, c, d does not affect q or the associated three-manifold. Second, once c and d are given, a and b are uniquely determined by $ad - bc = 1$ up to shifts of

the form $(a,b) \to (a,b) + t(c,d)$, $t \in \mathbb{Z}$. Under this transformation, q as defined in eqn. (2.12) is invariant. So the possible three-manifolds really only depend on the choice of the pair c, d of relatively prime integers, up to sign. For each such pair, we find integers a, b such that $ad - bc = 1$, and identify q via eqn. (2.12). This gives a manifold that we will call $M_{c,d}$. This family of manifolds were first discussed in the context of three-dimensional gravity in [9].

The Geometry of $M_{c,d}$

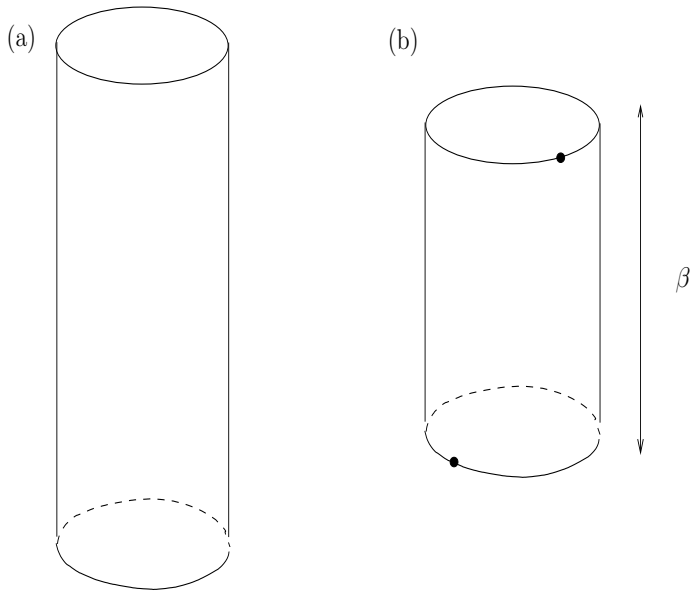


Figure 1: a) An infinite cylinder representing AdS_3. The boundary of the cylinder represents conformal infinity; time translations act by vertical shifts. b) A slice of height determined by β. The manifold $M_{0,1}$ is built by gluing together the top and bottom, after a rotation that identifies the boundary points marked by solid dots.

The simplest such manifold is $M_{0,1}$, which we will now describe in more detail. If we identify the real one-parameter subgroup $\mathrm{diag}(e^b, e^{-b})$ of $SL(2,\mathbb{C})$ as the group of time translations, the AdS_3 metric (2.1) can be put in the form

$$ds^2 = \cosh^2 r \; dt^2 + dr^2 + \sinh^2 r \; d\phi^2, \tag{2.13}$$

with $-\infty < t < \infty$, $0 \le r < \infty$, and $0 \le \phi \le 2\pi$. Described in this coordinate system is the subset $AdS_3' \subset AdS_3$ on which Γ acts nicely; its topology is $D \times \mathbb{R}$, where D is a

two-dimensional open disc parameterized by r and ϕ. Conformal infinity is at $r = \infty$. (See Fig. 1.) The element $\mathrm{diag}(e^b, e^{-b})$ acts by $t \to t + b$. The group of spatial rotations is the one-parameter group $\mathrm{diag}(e^{i\theta}, e^{-i\theta})$, acting by $\phi \to \phi + \theta$.

The group element W therefore generates a combined time-translation and spatial rotation. To explicitly divide by W, we "cut" AdS_3' at times $t = 0$ and $t = 2\pi \,\mathrm{Im}\,\tau$. Then we glue together the top and bottom of the region $0 \le t \le 2\pi \,\mathrm{Im}\,\tau$ after making a spatial rotation by an angle $2\pi \,\mathrm{Re}\,\tau$. This is sketched in Fig. 1b. The resulting spacetime $M_{0,1}$ is topologically $D \times S^1$, so its fundamental group is indeed $\mathbb{Z}$.

The path integral in this spacetime has a simple semiclassical meaning, since it may be interpreted in terms of Hamiltonian time evolution. A state is prepared at time zero and propagates a distance $\beta = 2\pi \,\mathrm{Im}\,\tau$ forward in Euclidean time. In this process, the state vector is multiplied by the time evolution operator $\exp(-\beta H)$, where H is the Hamiltonian. Then, after a spatial rotation by an angle $\theta = 2\pi \,\mathrm{Re}\,\tau$, which acts on the state by $\exp(-i\theta J)$, we glue the top and bottom of the figure, which results in taking the inner product of the final state with the initial state. The whole operation gives the trace $\mathrm{Tr}\,\exp(-\beta H - i\theta J)$ defined in the Hilbert space of perturbative fluctuations around AdS_3. This Hamiltonian interpretation of the path integral of $M_{0,1}$ will be the basis for evaluating it in Sec. 2.2.

The other manifolds $M_{c,d}$ are obtained from $M_{0,1}$ by modular transformations, that is, by diffeomorphisms that act non-trivially on the homology of Σ. This fact will allow us to evaluate their contributions to the path integral. These other manifolds may be thought of as Euclidean black holes.

For example, in the (t, ϕ) coordinates introduced above, $M_{0,1}$ involves the identifications

$$\phi + it \sim \phi + it + 2\pi \sim \phi + it + 2\pi\tau. \tag{2.14}$$

From the metric (2.13), note that the ϕ circle is contractible in the bulk, since the coefficient of $d\phi^2$ becomes zero at the origin $r = 0$. The manifold $M_{1,0}$ is described by the same coordinate system (2.13), but with new identifications

$$\phi + it \sim \phi + it + 2\pi \sim \phi + it - \frac{2\pi}{\tau}. \tag{2.15}$$

These identifications (2.15) may be written in the form of (2.14) by taking $\phi + it \to \frac{1}{\tau}(\phi + it)$. After this transformation, the coordinate that is contractible in the bulk now involves a combination of ϕ and t.

If we take τ to be imaginary, the scaling by $1/\tau$ has the effect of exchanging (and rescaling) ϕ and t. $M_{1,0}$ is hence a Euclidean black hole, in which the "time" circle is contractible rather than the "space" circle. If we rotate to Lorentzian signature by $t \to it$, then the locus $r = 0$ where the coefficient of dt^2 vanishes is the horizon of the black hole. In fact, $M_{1,0}$ is just the Euclidean version of the BTZ black hole [16]. The more general manifolds $M_{c,d}$ are often referred to as the $SL(2,\mathbb{Z})$ family of black holes.

Orbifolds

Finally, let us consider the extension in which Γ has a second generator given by (2.9). We may assume that $n = 2m$ is even, since the element $\mathrm{diag}(-1,-1)$ (the non-trivial element of the center of $SL(2,\mathbb{C})$) acts trivially on $\mathbb{CP}^1$. We also may as well assume that $m > 1$, since the case $m = 1$ leads to nothing new.

What we get when Γ has an additional generator with $m > 1$ is simply a three-dimensional space of the form $M_{c,d}/\mathbb{Z}_m$. The conformal boundary is still a Riemann surface of genus 1, and by changing q we can adjust its modular parameter as we wish. However, for $m > 1$, the group element (2.9) acts on AdS_3' with fixed points, meaning that $M_{c,d}/\mathbb{Z}_m$ has orbifold singularities. The fixed points are of codimension 2 and the singularities look locally like $\mathbb{R}^2/\mathbb{Z}_m$, where $\mathbb{Z}_m$ acts as a rotation by an angle $2\pi/m$. This produces a deficit angle

$$\theta = 2\pi(1 - 1/m). \tag{2.16}$$

The picture is sketched in Fig. 2.

At the classical level, the physical meaning in three-dimensional gravity of a codimension two singularity characterized by a deficit angle is usually that it represents the orbit of a massive particle. The mass of the particle is related to the deficit angle [17]. The particular values of the deficit angle in eqn. (2.16) are special because they are related to orbifolds.

Of course, there are many consistent theories of three-dimensional quantum gravity plus matter, obtained from various widely studied string theory and M-theory constructions. They each have massive particles of various sorts, the precise masses and spins being model-dependent. It would be hopeless to try to completely solve such a general theory.

The question of interest in the present paper is whether there exists a theory that is in some sense minimal and solvable. Our hypothesis is that the minimal theory should be described by smooth geometries without orbifold or deficit angle singularities. Pragmatically, it appears difficult to avoid the problem found in Sec. 3 simply by including a small

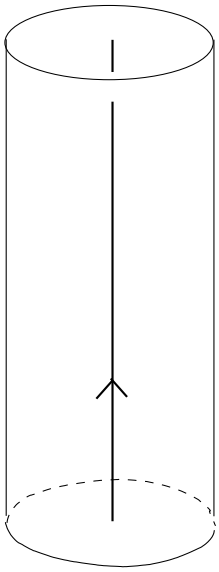


Figure 2: The black line represents an orbifold singularity in the interior of a spacetime that at infinity looks like AdS_3.

set of such singularities. In any event, were we to allow singularities, we would not know which ones to allow.

In the dual CFT, massive particles correspond to primary operators of positive dimension. Any two-dimensional CFT – and they are abundant, of course – can be interpreted as a dual AdS_3 theory. If this dual theory has a macroscopic, semiclassical interpretation, it can be interpreted in terms of AdS_3 with particles and black holes. From the holographic point of view, it is difficult to understand which CFTs have a semiclassical interpretation, but it is clear that, whether or not there is what one might call a minimal theory, there are many consistent theories with massive particles.

2.2. Evaluation Of The Partition Function

We write $Z_{c,d}(\tau)$ for the contribution to the partition function of the manifold $M_{c,d}$. Because the manifolds $M_{c,d}$ are all diffeomorphic to each other, the functions $Z_{c,d}(\tau)$ can all be expressed in terms of any one of them, say $Z_{0,1}(\tau)$, by a modular transformation. The formula is simply

$$Z_{c,d}(\tau) = Z_{0,1}((a\tau + b)/(c\tau + d)), \tag{2.17}$$

where a and b are any integers such that $ad-bc=1$. The partition function, or rather the sum of known contributions to it, is

$$Z(\tau)=\sum_{c,d}Z_{c,d}(\tau)=\sum_{c,d}Z_{0,1}((a\tau+b)/(c\tau+d)). \tag{2.18}$$

The summation here is over all integers c and d which are relatively prime and have $c\geq 0$.

This formula shows that the key point is to evaluate $Z_{0,1}(\tau)$. We recall that this contribution is simply Tr $\exp(-\beta H-i\theta J)$, computed in the Hilbert space that describes small fluctuations about AdS_3 (as opposed to black holes). If we know the eigenvalues of the commuting operators H and J in the Hilbert space of small fluctuations, then we can compute the trace.

In the most naive semiclassical approximation, $Z_{0,1}(\tau)$ is just $\exp(-I)$, where I is the classical action. In computing this action, one can not just evaluate the action (1.1) for the solution (2.13); such a computation would give an infinite answer, coming from the boundary at $r\to\infty$. The full action includes the Gibbons-Hawking boundary term, which has the opposite sign of the Einstein-Hilbert term (1.1). This extra term removes the divergence, and one arrives at a finite (negative) answer for the action of $M_{0,1}$ [16]:

$$I=-4\pi k\,\mathrm{Im}\,\tau \tag{2.19}$$

where $k=\ell/16G$. Therefore, in this approximation, we have

$$Z_{0,1}(\tau)\simeq|\bar{q}q|^{-k}. \tag{2.20}$$

The result (2.20) has a simple interpretation, in terms of facts explained in [6]. Three-dimensional pure gravity with the Einstein-Hilbert action (1.1) is dual to a conformal field theory with central charge $c_L=c_R=3\ell/2G=24k$. The formula $c_L=c_R=3\ell/2G$ is actually the semiclassical approximation of [2]. Depending on how the theory is regularized, there may be quantum corrections to this formula, but they preserve $c_L=c_R$ because the Einstein-Hilbert theory is parity symmetric. (By adding to the action a gravitational Chern-Simons term, we could generalize to $c_L\neq c_R$. We omit this here.) For our purposes, we simply parametrize the theory in terms of $k=c_L/24=c_R/24$. Since c_L and c_R are physical observables (defined in terms of the two-point function of the stress tensor in the boundary CFT, or an equivalent bulk computation), the theory when parametrized in this way does not depend on any choice of formalism. (By contrast, if we describe

the results in terms of the microscopic variable ℓ/G, then it is also necessary to describe the regularization.) It is conjectured in [13] that pure three-dimensional quantum gravity exists, if at all, only for integer k. However, for our present purposes, we do not need to know if this is correct.

Let L_0 and $\widetilde{L}_0$ be the Hamiltonians for left- and right-moving modes of the CFT. They are related to what we have called H and J by

$$\begin{aligned} H &= L_0 + \widetilde{L}_0 \\ J &= L_0 - \widetilde{L}_0. \end{aligned} \tag{2.21}$$

The CFT ground state has $L_0 = -c_L/24$, $\widetilde{L}_0 = -c_R/24$, or in the present context $L_0 = \widetilde{L}_0 = -k$. Equivalently, this state has $H = -2k$, $J = 0$. Its contribution to $\mathrm{Tr}\,\exp\left(-2\pi(\mathrm{Im}\,\tau)H + 2\pi i(\mathrm{Re}\,\tau)J\right)$ is $\exp(4\pi k \mathrm{Im}\,\tau) = |\bar{q}q|^{-k}$, as in eqn. (2.20).

Thus, although seemingly only a naive approximation, eqn. (2.20) actually gives the exact result for the contribution of the ground state to the partition function (if the theory is parametrized in terms of the central charges). As a result, it is also exact for $\mathrm{Im}\,\tau \to \infty$.

To get a complete answer, we also need to know the energies of excited states. In more than $2+1$ dimensions, the trace would receive contributions from a gas of gravitons as well as from other particles, if present. Their energies would in general receive complicated perturbative (and perhaps nonperturbative) corrections. Hence, although the function $Z_{0,1}(\tau)$ has a natural analog in any dimension, above $2+1$ dimensions one would not expect to be able to compute it precisely.

In 2+1 dimensions, there are no gravitational waves. Naively, there are no perturbative excitations at all above the AdS_3 vacuum, and hence one might at first expect the formula (2.20) to be exact. However, this is mistaken, because of the insight of Brown and Henneaux [2]. There must at least be states that correspond to Virasoro descendants of the identity, or in other words states obtained by repeatedly acting on the CFT vacuum $|\Omega\rangle$ with the stress tensor. If L_n and $\widetilde{L}_n$ are the left- and right-moving modes of the Virasoro algebra, then a general such state is

$$\prod_{n=2}^{\infty} L_{-n}^{u_n} \prod_{m=2}^{\infty} \widetilde{L}_{-m}^{v_m} |\Omega\rangle, \tag{2.22}$$

with non-negative integers u_n, v_m. (For a state of finite energy, almost all of these integers must vanish. The products begin with $n, m = 2$ since L_{-1} and $\widetilde{L}_{-1}$ annihilate the CFT

ground state.) A state of this form is an eigenstate of L_0 and $\widetilde{L}_0$ with $L_0 = -k + \sum_{n=2}^{\infty} n u_n$, $\widetilde{L}_0 = -k + \sum_{m=2}^{\infty} m v_m$. The contribution of these states to the partition function is then

$$Z_{0,1}(\tau) = |\bar{q}q|^{-k} \frac{1}{\prod_{n=2}^{\infty} |1-q^n|^2}. \tag{2.23}$$

It is convenient to introduce the Dedekind η function, defined by

$$\eta(\tau) = q^{1/24} \prod_{n=1}^{\infty} (1-q^n). \tag{2.24}$$

Eqn. (2.23) can then be rewritten

$$Z_{0,1}(\tau) = \frac{1}{|\eta(\tau)|^2} |\bar{q}q|^{-(k-1/24)} |1-q|^2. \tag{2.25}$$

This formula will be useful in Sec. 3 because $(\mathrm{Im}\,\tau)^{1/2}|\eta(\tau)|^2$ is modular-invariant.

We will show shortly how the Virasoro descendants of eqn. (2.22) arise in an analysis along the lines of Brown and Henneaux. The argument will also show that the formula (2.23) is exact to all order of perturbation theory. As for whether there are nonperturbative corrections to eqn. (2.23), it is not clear whether this question is well-defined, since it may be impossible to separate the question of unknown nonperturbative corrections to $Z_{0,1}$ from the more general question of unknown nonperturbative corrections to the exact partition function Z. At any rate, it will be easier to discuss what nonperturbative corrections would mean after explaining why eqn. (2.23) agrees with perturbation theory.

Before presenting this argument, let us discuss what the answer means from the general point of view of quantum mechanical perturbation theory. Usually, one defines an effective action $I_{\rm eff}$ such that the partition function is $Z = \exp(-kI_{\rm eff})$. $I_{\rm eff}$ is equal to the classical action I plus contributions generated from r-loop diagrams:

$$I_{\rm eff} = I + \sum_{r=1}^{\infty} k^{-r} I_r. \tag{2.26}$$

Here I_r is a function generated from Feynman diagrams with r loops. In a general theory, the expansion (2.26) is only asymptotic in $1/k$; there may be further contributions that are non-perturbatively small for $k \to \infty$. However, in the present context, the formula (2.23) implies that (for perturbation theory around the manifold $M_{0,1}$) there are no corrections beyond one-loop order; $I_{\rm eff}$ is simply the classical action plus a one-loop correction. In other

words, the formula implies that if $Z_{0,1}$ is directly computed in perturbation theory, then the perturbation series terminates with the one-loop term. (Such a direct computation is also briefly discussed below.)

In a sense, this should not come as a surprise. The gauge theory interpretation of three-dimensional gravity [18,19] (or even more naively the absence of any local excitations) suggests that in some sense it is an integrable system. It is often true for quantum integrable systems that properly chosen quantities are one-loop exact.

Derivation Of The Formula

In most of physics, the quantum Hilbert space of a theory is constructed by quantizing an appropriate classical phase space. This framework may ultimately be inadequate for quantum gravity, but it is certainly adequate for constructing the perturbative Hilbert space that is needed to compute $Z_{0,1}$ in perturbation theory.

The phase space of any physical theory is simply the space of its classical solutions. In $2+1$-dimensional gravity with AdS boundary conditions at infinity, naively speaking the phase space relevant to perturbation theory consists of only a single point, since any solution that is close to the AdS_3 solution (close enough to exclude black holes) actually is diffeomorphic to AdS_3.

However, we must be more careful here. In General Relativity, one should divide only by diffeomorphisms that approach the identity fast enough at infinity. After doing so, one constructs the classical phase space $\mathcal{M}$; it parametrizes classical solutions that obey the boundary conditions modulo diffeomorphisms that vanish fast enough at infinity. One is then left with an action of a group G that consists of those diffeomorphisms that preserve the boundary conditions, modulo those that vanish fast enough at infinity that they are required to act trivially on physical states.

In General Relativity in $3+1$ or more dimensions, in a spacetime that is asymptotically Lorentzian, the group G is the Poincaré group. This is why quantum gravity, in an asymptotically Lorentzian spacetime, is Poincaré invariant. In $2+1$-dimensional gravity with negative cosmological constant, the obvious analog of this answer would be $SL(2,\mathbb{R})\times SL(2,\mathbb{R})$ (or a group locally isomorphic to that one), the group of symmetries of AdS_3. However, Brown and Henneaux showed [2] that the actual answer turns out to be the infinite-dimensional group $G = \mathrm{diff}\, S^1 \times \mathrm{diff}\, S^1$ (which contains $SL(2,\mathbb{R})\times SL(2,\mathbb{R})$ as a subgroup). By computing Poisson brackets, they also showed that the quantum symmetry

group will be not G but a central extension with (in a semiclassical approximation) $c_L = c_R = 3\ell/2G$. We denote this central extension as $\widehat{\mathrm{diff}}\, S^1 \times \widehat{\mathrm{diff}}\, S^1$.

We will carry the analysis of [2] just slightly farther to describe the phase space $\mathcal{M}$ and the resulting energy levels. First of all, $\mathcal{M}$ is a homogeneous space for G, since if we divide by all diffeomorphisms, then the classical solutions that obey the boundary conditions (and are close enough to AdS_3) are equivalent to AdS_3. So $\mathcal{M}$ is G/H, where H is some subgroup of G. In fact, H is the subgroup of G that leaves fixed a given point on $\mathcal{M}$. Differently put, the symmetry group of a point on $\mathcal{M}$ is isomorphic to H. In the present context, we simply pick the point on $\mathcal{M}$ corresponding to AdS_3 and observe that its symmetry group is $SL(2,\mathbb{R}) \times SL(2,\mathbb{R})$. So this is H, and the phase space is $\mathcal{M} = (\mathrm{diff}\, S^1 \times \mathrm{diff}\, S^1)/(SL(2,\mathbb{R}) \times SL(2,\mathbb{R})) = (\mathrm{diff}\, S^1/SL(2,\mathbb{R}))^2$.

In general, the quantization of a homogeneous space G/H should be expected to give a Hilbert space that is an irreducible representation R of G (or more generally of a central extension of G). Moreover, as G/H contains an H-invariant point, the representation R will contain a vector that is an eigenvector for the action of H (and hence is H-invariant if H is a simple non-abelian group).

In the present problem, a representation of the Virasoro group $\widehat{\mathrm{diff}}\, S^1$ in which L_0 is bounded below and in which there is an $SL(2,\mathbb{R})$-invariant vector is uniquely determined once the central charge $c = 24k$ is given. It is the "vacuum" representation, containing a vector $|\Omega\rangle$ with $(L_n + k\delta_{n,0})|\Omega\rangle = 0$, $n \geq -1$. This representation is spanned by states of the form

$$\prod_{n=2}^{\infty} L_{-n}^{a_n} |\Omega\rangle, \tag{2.27}$$

with energy

$$\epsilon = -k + \sum_{n=2}^{\infty} n a_n. \tag{2.28}$$

Quantizing $(\mathrm{diff}\, S^1/SL(2,\mathbb{R}))^2$ gives a tensor product of two such representations for the two factors of $G = (\mathrm{diff}\, S^1)^2$, and in this way we arrive at the spectrum claimed in eqn. (2.22), and hence at the formula (2.23) for the partition function.

Quantization of homogeneous spaces of $\mathrm{diff}\, S^1$ is described in [20]. The fact that quantization of the quotient $\mathrm{diff}\, S^1/SL(2,\mathbb{R})$ gives the vacuum representation of $\widehat{\mathrm{diff}}\, S^1$ is explained from several points of view in eqns. (170), (172), and (174) of that paper.

The states that we have described can be regarded as boundary excitations, supported near the boundary of AdS_3, since $\mathcal{M}$ collapses to a point if we are allowed to make general

changes of coordinate near the boundary. We call these boundary excitations the Brown-Henneaux (BH) states, since they are so closely related to the analysis in [2].

Corrections?

Now let us ask to what extent there may be corrections to the spectrum just described.

First we consider the question of quantum corrections to the formula (2.28) for this family of states. We can dispose of this question immediately: this particular family of states transforms in an irreducible representation of the symmetry group, and this representation has no possible deformations once the central charge is fixed.

So there are no quantum corrections to the energies of the states just described. Could there be additional states contributing to $Z_{0,1}$?

In general, in three-dimensional gravity, there may be additional states of various kinds, corresponding for instance to massive or even massless particles. What do we expect in "pure" three-dimensional gravity, if it exists? It is important to distinguish two length scales: the AdS_3 length scale ℓ, and the Planck length G. In the semiclassical regime of $k \to \infty$, we have $G << \ell$.

In eqn. (2.28), energies are measured in units of $1/\ell$. A reasonable minimum condition for anything that one might call a theory of pure gravity is that the energy of any state that is not part of the BH spectrum goes to infinity in the semiclassical limit $k \to \infty$ (when measured relative to the energy of the ground state). States that are not part of the BH spectrum and whose excitation energies remain fixed for $k \to \infty$ would be interpreted in terms of non-gravitational fields propagating in AdS_3.

As for what sort of states might have energies that go to infinity for $k \to \infty$, we certainly expect BTZ black holes, with energy bounded below by k. As for what else there may be, the sky is the limit in terms of conceivable speculations. If so inclined, one can postulate solitons with excitation energies proportional to k, states similar to D-branes with energies proportional to $k^{1/2}$, etc. Orbifold singularities with fixed deficit angle, as described in relation to eqn. (2.16), would also have excitation energy of order k.

Our point of view is that the question of whether there are nonperturbative corrections to $Z_{0,1}$ from states whose excitation energy diverges as $k \to \infty$ is hard to separate from the more general question of unknown nonperturbative corrections to the exact partition function Z. Since the perturbative evaluation of $Z_{0,1}$ gives a convergent and physically sensible result (which is even one-loop exact), we may as well regard hypothetical contributions in which $M_{0,1}$ is enriched with a soliton, a D-brane, an orbifold singularity, or some

other unknown type of excitation as representing different, presently unknown sectors of the path integral. In this sense, the formula for $Z_{0,1}$ is exact.

Comparison To Perturbation Theory

Since we claim that the formula for $Z_{0,1}$ is one-loop exact, the question arises of why not to simply calculate it by evaluating the relevant one-loop determinants.

We know of two efforts to do so. In [16], a formula much simpler than our result (2.23) is claimed. The analysis relies on the relation of the relevant product of determinants to Ray-Singer and Reidemeister torsion [11]. For an irreducible flat connection on a compact manifold without boundary, the torsion is simply a number, but for a manifold with boundary (such as $M_{0,1}$ effectively is), the torsion must be understood as measure on a certain moduli space associated with the boundary. A rather subtle treatment of this is needed, we suspect, to compute the one-loop correction using its interpretation via torsion.

On the other hand, in [21], the one-loop correction was studied by expressing the determinants in terms of the appropriate heat kernels, which were evaluated using a method of images (analogous to the derivation of the Selberg trace formula). We believe that this method is conceptually completely correct. The claimed result is qualitatively similar to our formula (2.23) (for example, it has an expansion in integer powers of q and $\bar{q}$), and we hope that it will prove possible to resolve any discrepancies between the result in [21] and the one claimed here.

As for why it is much simpler to compute the one-loop correction via the Hamiltonian route that we have followed, this should not really come as a surprise. In general, path integrals on a product $S^1 \times Y$ are often most easily evaluated by constructing an appropriate Hilbert space in quantization on Y and then taking a trace.[3] That is especially likely to be true in the present situation, in which the relevant excitations are rather subtle boundary excitations, already known but difficult to rediscover.

[3] An excellent example of this statement is given by Chern-Simons gauge theory with a compact gauge group for the case that Y is a compact Riemann surface without boundary. The partition function on $S^1 \times Y$ is an integer, the dimension of the physical Hilbert space $\mathcal{H}$ associated with Y. It is not too hard to describe $\mathcal{H}$, compute its dimension, and thereby learn the value of the path integral on $S^1 \times Y$. But direct evaluation of this path integral by Lorentz-covariant methods is difficult, even in perturbation theory.

2.3. The General Form Of The Partition Function

We conclude with an analysis of the following question: in general, what sort of function can be written as $\mathrm{Tr}\,\exp(-\beta H)$, where H is a hermitian operator on a Hilbert space $\mathcal{H}$? We assume that H is constrained so that $\mathrm{Tr}\,\exp(-\beta H)$ is convergent whenever $\mathrm{Re}\,\beta > 0$. (According to standard assumptions about pure three-dimensional gravity, this condition is satisfied, but we discuss in Sec. 4.1 one way that it might fail.)

For $\mathrm{Tr}\,\exp(-\beta H)$ to be convergent, the number n_E of states of energy less than E must be finite for each E. Indeed, the trace is bounded below by $n_E \exp(-\beta E)$. Finiteness of n_E for all E implies in particular that H must have a discrete spectrum.

Let E_* be any value of the energy and let $E_1, \dots, E_n$ be the energy eigenvalues that are no greater than E_*. Their contribution to the partition function is $\sum_{i=1}^n \exp(-\beta E_i)$, and the full partition function is therefore

$$\mathrm{Tr}\,\exp(-\beta H) = \sum_{i=1}^{n} \exp(-\beta E_i) + \mathcal{O}(\exp(-\beta E_*)). \tag{2.29}$$

Here $\mathcal{O}(\exp(-\beta E_*))$ is a function that is bounded by a multiple of $\exp(-\beta E_*)$.

We have to be careful here with one point: any finite sum of states of energy greater than E_* makes a contribution of the form $\mathcal{O}(\exp(-\beta E_*))$, but in general an infinite set of states with energy $\geq E_*$ may make a contribution that is not bounded in this way. However, our hypothesis that $\mathrm{Tr}\,\exp(-\beta H)$ converges whenever $\mathrm{Re}\,\beta > 0$ ensures that the number n_E of states of energy less than E grows with E more slowly than any exponential. This is enough to justify the error estimate in eqn. (2.29).

Our problem is slightly more general. We have a pair of commuting hermitian operators H and J, and we are computing the trace $\mathrm{Tr}\,\exp(-\beta H - i\theta J)$. However, the eigenvalues of J are integers, so

$$\mathrm{Tr}\,\exp(-\beta H - i\theta J) = \sum_{j\in\mathbb{Z}} e^{-ij\theta}\,\mathrm{Tr}_j\,\exp(-\beta H), \tag{2.30}$$

where Tr_j is a trace in the subspace $\mathcal{H}_j$ in which J acts with eigenvalue j. The functions $\mathrm{Tr}_j\exp(-\beta H)$ should be constrained exactly as in eqn. (2.29).

In Sec. 3, we will find instead that the evaluation of known contributions to $\mathrm{Tr}_0\exp(-\beta H)$ take the form just described up to a certain energy E_*, beyond which this form breaks down. The result looks like

$$\sum_{i=1}^{n} \exp(-\beta E_i) + \exp(-\beta E_*)(f + \sum_{s=1}^{\infty} f_s \beta^{-s}) \tag{2.31}$$

where f is a negative integer, rather than being positive as it should be, and the corrections decay as power laws, rather than exponentials. (The series in (2.31) may be only asymptotic.) Both the fact that f is negative and the fact that there are power-law corrections mean that this function cannot be interpreted as Tr $\exp(-\beta H)$.

Finally, let us compare the assertion that Tr $\exp(-\beta H)$ can only converge if H has a discrete spectrum with standard results about the thermodynamics of physically sensible systems with continuous spectrum, such as a free gas of particles on the real line. What is usually computed in such a case is the free energy per unit volume; the total free energy is infinite simply because the volume is infinite. To make the total free energy finite, one can place the gas in a finite volume; then the spectrum is discrete and Tr $\exp(-\beta H)$ converges. A gas of particles in Anti-de Sitter space will have a discrete spectrum and finite partition function Tr $\exp(-\beta H)$ as well. This is because the motion of any finite energy particle in Anti-de Sitter space is restricted to a finite volume region in the interior of AdS.

3. Computing the Sum over Geometries

As explained in Secs. 1 and 2, the known contributions to the partition function of pure gravity in a spacetime asymptotic to AdS_3 come from smooth geometries $M_{c,d}$, where c and d are a pair of relatively prime integers (with a pair c, d identified with $-c, -d$). Their contribution to the partition function, including the contribution from the Brown-Henneaux excitations, is

$$Z(\tau) = \sum_{c,d} Z_{0,1}(\gamma\tau), \tag{3.1}$$

where

$$\gamma\tau = \frac{a\tau + b}{c\tau + d}, \qquad \gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in SL(2,\mathbb{Z}) \tag{3.2}$$

and

$$Z_{0,1}(\tau) = \left| q^{-k} \prod_{n=2}^{\infty} (1-q^n)^{-1} \right|^2 = \frac{|\bar{q}q|^{-k+1/24}|1-q|^2}{|\eta(\tau)|^2}. \tag{3.3}$$

The summation in (3.1) is over all relatively prime c and d with $c \geq 0$. Since $Z_{0,1}(\tau)$ is invariant under $\tau \to \tau + 1$, the summand in (3.1) is independent of the choice of a and b in (3.2). This sum over c and d in (3.1) should be thought of as a sum over the coset $PSL(2,\mathbb{Z})/\mathbb{Z}$, where $\mathbb{Z}$ is the subgroup of $PSL(2,\mathbb{Z}) = SL(2,\mathbb{Z})/\{\pm 1\}$ that acts by

$\tau \to \tau + n$, $n \in \mathbb{Z}$. Given any function of τ, such as $Z_{0,1}(\tau)$, that is invariant under $\tau \to \tau + 1$, one may form a sum such as (3.1), known as a Poincaré series.

The function $\sqrt{\mathrm{Im}\,\tau}|\eta(\tau)|^2$ is modular-invariant. We can therefore write $Z(\tau)$ as a much simpler-looking Poincaré series,

$$Z(\tau) = \frac{1}{\sqrt{\mathrm{Im}\,\tau}|\eta(\tau)|^2} \sum_{c,d} \left(\sqrt{\mathrm{Im}\,\tau}|\bar{q}q|^{-k+1/24}|1-q|^2\right)\Big|_\gamma , \tag{3.4}$$

where $(\ldots)|_\gamma$ is the transform of an expression $(\ldots)$ by γ. Writing out explicitly $|1-q|^2 = 1 - q - \bar{q} + \bar{q}q$, we see that we really need a sum of four Poincaré series, each of the form

$$E(\tau; n, m) = \sum_{c,d} \left(\sqrt{\mathrm{Im}\,\tau} q^{-n} \bar{q}^{-m}\right)\Big|_\gamma , \tag{3.5}$$

with $n - m$ equal to 0 or ± 1. Precisely this sum, or rather its s-dependent generalization introduced below, has been studied in Sec. 3.4 of [22], as was pointed out by P. Sarnak.[4] If we set $\kappa = n + m$, $\mu = m - n$, and use the fact that $\mathrm{Im}\,(\gamma\tau) = \mathrm{Im}\,\tau/|c\tau + d|^2$, then the basic Poincaré series can be written

$$E(\tau; \kappa, \mu) = \sqrt{\mathrm{Im}\,\tau} \sum_{c,d} |c\tau + d|^{-1} \exp\left\{2\pi\kappa\ \mathrm{Im}\ \gamma\tau + 2\pi i\mu\ \mathrm{Re}\ \gamma\tau\right\}. \tag{3.6}$$

When $\kappa = 0$ and $\mu = 0$, this sum is a non-holomorphic Eiseinstein series of weight $1/2$. Sometimes we omit τ and write just $E(\kappa, \mu)$. In terms of this function, the partition function is

$$\begin{aligned} Z(\tau) = \frac{1}{\sqrt{\mathrm{Im}\,\tau}|\eta(\tau)|^2} \big(&E(2k - 1/12, 0) + E(2k + 2 - 1/12, 0) \\ &- E(2k + 1 - 1/12, 1) - E(2k + 1 - 1/12, -1)\big). \end{aligned} \tag{3.7}$$

3.1. The Regularized Sum

The sums (3.6) that define the Poincaré series we need are divergent. (Such divergences have been encountered before in similar sums related to three-dimensional gravity [10,23,14].) Using

$$\frac{a\tau + b}{c\tau + d} = \frac{a}{c} - \frac{1}{c(c\tau + d)}, \tag{3.8}$$

[4] Actually, the generalization to arbitrary integer values of $n - m$ is considered in [22]. We would encounter the same generalization if we modify the original Einstein-Hilbert action (1.1) to include the gravitational Chern-Simons term.

one may show that if $\tau = x + iy$ then

$$\begin{aligned} \mathrm{Im}\,(\gamma\tau) &= \frac{y}{(cx+d)^2 + c^2y^2} \\ \mathrm{Re}\,(\gamma\tau) &= \frac{a}{c} - \frac{cx+d}{c((cx+d)^2+c^2y^2)}. \end{aligned} \tag{3.9}$$

For $\mu = 0$, the exponential factor in the definition of E is

$$\exp\{2\pi\kappa\ \mathrm{Im}\ \gamma\tau\} = \exp\left\{2\pi\kappa\ \frac{y}{(cx+d)^2+c^2y^2}\right\}, \tag{3.10}$$

and this goes to 1 for $c, d \to \infty$. So the first two terms in (3.7) diverge linearly as $\sum_{c,d} |c\tau + d|^{-1}$ at large c and d. The other terms also diverge, though more slowly.

We claim that this divergence has a natural regularization. On the upper half plane, which we call $\mathcal{H}$, there is a natural $SL(2,\mathbb{R})$-invariant Laplacian:

$$\Delta = -y^2\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right). \tag{3.11}$$

A short calculation shows that the function $y^{1/2}$ is an eigenfunction of Δ: $\Delta(y^{1/2}) = (1/4)y^{1/2}$. Since Δ is $SL(2,\mathbb{R})$-invariant, the same is true of $(\mathrm{Im}\,\gamma\tau)^{1/2}$ for any $\gamma \in SL(2,\mathbb{Z})$ (or even $SL(2,\mathbb{R})$):

$$\Delta\sqrt{\mathrm{Im}\,(\gamma\tau)} = \frac{1}{4}\sqrt{\mathrm{Im}\,(\gamma\tau)}. \tag{3.12}$$

Using this, one may verify that although the Poincaré series for $E(\tau;\kappa,\mu)$ is divergent, the corresponding series for $(\Delta - 1/4)E(\tau;\kappa,\mu)$ actually converges. (This series is obtained by acting termwise with $(\Delta - 1/4)$ on the Poincaré series for E.) To see this, we just observe, using (3.10), that contributions in which derivatives appearing in Δ act on the exponential factor in (3.6) get an extra convergence factor of order $1/(\max(c,d))^2$ and lead to a convergent sum over c and d. On the other hand, contributions in which none of the derivatives act on the exponential actually vanish, because of (3.12).

So $(\Delta - 1/4)E$ requires no regularization. Similarly, if we set $F = \sqrt{\mathrm{Im}\,\tau}|\eta|^2$, then $(\Delta - 1/4)(FZ)$ requires no regularization, where Z is the partition function. Since no regularization is required, we assume that the naive sum gives $(\Delta - 1/4)(FZ)$ correctly.

The Laplacian Δ, acting on the Hilbert space of square-integrable $SL(2,\mathbb{Z})$-invariant functions on $\mathcal{H}$, has a continuous spectrum starting at $1/4$. It also has a discrete spectrum, but there are no discrete modes with eigenvalue $1/4$. So the operator $\Delta - 1/4$ is invertible

acting on square-integrable functions, and roughly speaking, we now want to argue that once $(\Delta - 1/4)(FZ)$ is known, FZ is uniquely determined.

There is a subtlety in making this argument, because actually Z and FZ are not themselves square-integrable. Rather, we expect that Z grows exponentially for $y \to \infty$, because the AdS_3 vacuum has negative energy $-2k$. There are also Brown-Henneaux excitations of this vacuum, again with negative energy, and they contribute additional exponentially growing terms for $y \to \infty$. If we assume that all exponentially growing contributions to Z come from the known Brown-Henneaux states, then this, together with a knowledge of $(\Delta - 1/4)(FZ)$, is enough to determine Z uniquely.

But for our main application, we do not really need this assumption (which appears to be not quite correct according to the analysis of Sec. 3.3). The explicit calculation of Secs. 3.2 and 3.3 will give us a function Z, compatible with the Poincaré series for $(\Delta-1/4)(FZ)$, such that the fastest growing exponentials agree with the Brown-Henneaux spectrum. The leading departure from the contributions of that spectrum are determined at the end of Sec. 3.3 and take the form

$$Z' = \frac{1}{|\eta|^2}\Big(-6 + \frac{(\pi^3 - 6\pi)(11+24k)}{9\zeta(3)} y^{-1} + \frac{5(53\pi^6 - 882\pi^2) + 528(\pi^6 - 90\pi^2)k + 576(\pi^6 - 90\pi^2)k^2}{2430\zeta(5)} y^{-2} + \mathcal{O}(y^{-3})\Big). \tag{3.13}$$

This function cannot be written as a sum of exponentials with positive integer coefficients, since the leading coefficient is negative, and the corrections do not have the right form. Any other candidate for Z would be obtained by adding a correction $\widetilde{Z}$ such that $(\Delta - 1/4)(F\widetilde{Z}) = 0$. But that equation is not obeyed by any function $\widetilde{Z}$ that would solve our problem; the equation $(\Delta - 1/4)(F\widetilde{Z}) = 0$ is not obeyed by $-Z'$ or by any function that differs from it for $y \to \infty$ by a sum of exponentials with positive integer coefficients. So the problem that we will find is not affected by adding to the function Z that we compute an additional contribution $\widetilde{Z}$ that obeys $(\Delta - 1/4)(F\widetilde{Z}) = 0$.

More Convenient Alternative

Although this argument is satisfying conceptually, it does not give a convenient way to determine Z in practice. A much more convenient method is to adapt ζ-function regularization [11] to this problem.

In the present context, the analog of ζ-function regularization is to replace the Poincaré series (3.5) by a more general one depending on a parameter s:

$$E(\tau; s, n, m) = \sum_{c,d} \left((\mathrm{Im}\,\tau)^s q^{-n} \bar{q}^{-m} \right)\big|_\gamma \,. \tag{3.14}$$

Since $(\mathrm{Im}\,\gamma\tau)^s = y^s/|c\tau + d|^{2s}$, the series $E(\tau; s, n, m)$ converges for $\mathrm{Re}\,s > 1$. Our original problem concerns the case $s = 1/2$. As we will see shortly, $E(\tau; s, n, m)$, defined initially for $\mathrm{Re}\,s > 1$, can be analytically continued to $s = 1/2$ without any problem. This analytic continuation gives a natural way to define the original function $E(\tau; n, m)$ and therefore the partition function Z. This turns out to be a very practical and useful method to study $E(\tau; n, m)$.

The only problem with this approach is that the physical meaning of the parameter s in three-dimensional gravity is unclear; hence, it is not clear *a priori* that the analytic continuation will give the right answer. The argument involving $(\Delta - 1/4)(FZ)$ is clearer conceptually, because $(\Delta - 1/4)(FZ)$ is a physical observable and we simply use the fact that the formal path integral expression for it converges.

However, it is not difficult to show that the two methods give the same result. For this, we observe that the function $H(\tau; s, m, n) = (\Delta + s(s-1))E(\tau; s, m, n)$ can also be represented by a Poincaré series, convergent when $\mathrm{Re}\,s > 0$. For $\mathrm{Re}\,s > 1$, where both Poincaré series converge, the two functions E and H, both defined by their Poincaré series, obey $H = (\Delta + s(s-1))E$. This automatically remains true after analytic continuation to $s = 1/2$. So the function $E(\tau; 1/2, m, n)$ defined by analytic continuation from $\mathrm{Re}\,s > 1$ has the property that $(\Delta - 1/4)E(\tau; 1/2, m, n)$ is given by the obvious Poincaré series.

In terms of $\kappa = m + n$, $\mu = m - n$, the regularized Poincaré series is

$$E(s, \kappa, \mu) = \sum_{c,d} \frac{y^s}{|c\tau + d|^{2s}} \exp\left\{ 2\pi\kappa\ \mathrm{Im}\ \gamma\tau + 2\pi i\mu\ \mathrm{Re}\ \gamma\tau \right\}. \tag{3.15}$$

When $\kappa = 0$ and $\mu = 0$, this sum is the non-holomorphic Eisenstein series of weight s.

3.2. Poisson Resummation

Sec. 3.4 of [22] contains precisely what we need to analyze the sum (3.15), make the analytic continuation, and determine if the partition function Z is physically sensible.

First, define $d = d' + nc$, where n is an integer and d' runs from 0 and $c-1$. We may separate out the sum over n in (3.15) to get

$$E(s,\kappa,\mu) = y^s e^{2\pi(\kappa y + i\mu x)} + \sum_{c>0} \sum_{d' \in \mathbb{Z}/c\mathbb{Z}} \sum_{n\in\mathbb{Z}} f(c,d',n) \tag{3.16}$$

where

$$f(c,d',n) = \frac{y^s}{|c(\tau+n)+d'|^{2s}} \exp\left\{ \frac{2\pi\kappa y}{|c(\tau+n)+d'|^2} + 2\pi i\mu \left(\frac{a}{c} - \frac{cx+d}{c|c(\tau+n)+d'|^2} \right) \right\}. \tag{3.17}$$

The first term in eqn. (3.16) comes from $c = 0, d = 1$.

The Poisson summation formula allows us to turn the sum over n into a sum over a Fourier conjugate variable $\hat{n}$

$$\sum_{n\in\mathbb{Z}} f(c,d',n) = \sum_{\hat{n}\in\mathbb{Z}} \hat{f}(c,d',\hat{n}) \tag{3.18}$$

where $\hat{f}(c,d',\hat{n})$ is the Fourier transform

$$\begin{aligned} \hat{f}(c,d',\hat{n}) &= \int_{-\infty}^{\infty} dn\ e^{2\pi i n\hat{n}} f(c,d',n) \\ &= \exp\left(2\pi i \left(\frac{\mu a - \hat{n} d'}{c} - \hat{n}x \right)\right) \int_{-\infty}^{\infty} dt\ e^{2\pi i \hat{n} t} \left(\frac{y}{c^2(t^2+y^2)} \right)^s \exp\left\{ \frac{2\pi(\kappa\, y - i\mu t)}{c^2(t^2+y^2)} \right\}. \end{aligned} \tag{3.19}$$

We have written the integral in terms of a shifted integration variable $t = n + x + \frac{d'}{c}$. Upon Taylor expanding the exponential that appears in the integral and introducing $T = t/y$, we get

$$\begin{aligned} \hat{f}(c,d',\hat{n}) &= \sum_{m=0}^{\infty} c^{-2(s+m)} e^{2\pi i \left(\frac{\mu a - \hat{n} d'}{c} - \hat{n}x\right)} \frac{(2\pi)^m}{m!} \int_{-\infty}^{\infty} dt\ e^{2\pi i \hat{n} t} \left(\frac{y}{t^2+y^2} \right)^{m+s} \left(\kappa - i\mu \frac{t}{y} \right)^m \\ &= \sum_{m=0}^{\infty} c^{-2(s+m)} e^{2\pi i \left(\frac{\mu a - \hat{n} d'}{c} - \hat{n}x\right)} \frac{(2\pi)^m}{m!} y^{1-m-s} \int_{-\infty}^{\infty} dT\ e^{2\pi i \hat{n} T y} \left(1+T^2\right)^{-m-s} (\kappa - i\mu T)^m . \end{aligned} \tag{3.20}$$

In (3.20), c and d' do not appear in the integrals but only in the elementary prefactors, so we can study the sums over c and d' explicitly. Note that, as μ is an integer, (3.20) depends on a only modulo c. The value of a mod c is determined by d', given that $ad' = 1$ mod c. For a given d', such an a exists if and only if d' lies in the set $(\mathbb{Z}/c\mathbb{Z})^*$ of residue

classes mod c that are invertible multiplicatively. So, dropping the prime from d', we may write the sum over that variable as

$$S(-\hat{n},\mu;c) = \sum_{d\in(\mathbb{Z}/c\mathbb{Z})^*} \exp\left\{2\pi i\left(\frac{-\hat{n}d+\mu d^{-1}}{c}\right)\right\}, \tag{3.21}$$

where $d^{-1}\in(\mathbb{Z}/c\mathbb{Z})^*$ is the multiplicative inverse of d. This sum is known as a Kloosterman sum.

Rearranging the sums in (3.16), we now have

$$E(s,\kappa,\mu) = y^s e^{2\pi(\kappa y+i\mu x)} + \sum_{\hat{n}} e^{-2\pi i\hat{n}x} E_{\hat{n}}(s,\kappa,\mu) \tag{3.22}$$

where

$$E_{\hat{n}}(s,\kappa,\mu) = \sum_{m=0}^{\infty} I_{m,\hat{n}}(s,\kappa,\mu)\; y^{1-m-s}\;\left(\sum_{c=1}^{\infty} c^{-2(m+s)} S(-\hat{n},\mu;c)\right). \tag{3.23}$$

Here we have defined the integral

$$I_{m,\hat{n}}(s,\kappa,\mu) = \frac{(2\pi)^m}{m!}\int_{-\infty}^{\infty} dT\; e^{2\pi i\hat{n}Ty}\left(1+T^2\right)^{-m-s}(\kappa-i\mu T)^m\,. \tag{3.24}$$

Note that (3.23) is independent of x, so that (3.22) has the form of a Fourier expansion in x with Fourier coefficients $E_{\hat{n}}(s,\kappa,\mu)$ given by (3.23). These Fourier coefficients are typically complicated functions of y, since the integral (3.24) depends on y.

These Fourier coefficients are precisely what we want. In view of (3.7), the Fourier expansion of the function E with respect to x will give a similar expansion for the partition function Z. The Fourier coefficients of Z are the functions $\mathrm{Tr}_j\,\exp(-\beta H)$, the partition function restricted to states with angular momentum $J=j$. These are the functions that we want to understand.

We are almost ready to analyze what happens when we continue the formulas to $s=1/2$. For $m>0$, the integral in (3.24) is convergent for $\mathrm{Re}\,s>0$, and likewise the sum

$$\sum_{c=1}^{\infty} c^{-2(m+s)} S(-\hat{n},\mu;c) \tag{3.25}$$

converges for $\mathrm{Re}\,s>0$. A problem does occur for $m=0$, since then neither the integral nor the sum is convergent at $s=1/2$. We return to this shortly.

3.3. The $\hat{n} = 0$ Mode

Let us first consider the Fourier mode which is constant in x, i.e. the $\hat{n} = 0$ term in (3.22). In this case the integral (3.24) is independent of y, and may be evaluated explicitly. For $\mu = 0$, the result can be expressed in terms of Γ functions

$$I_{m,0}(s,\kappa,0) = \kappa^m \frac{2^m \pi^{m+1/2} \Gamma(s+m-1/2)}{m!\Gamma(s+m)}, \tag{3.26}$$

while for $\mu = \pm 1$, we require also hypergeometric functions

$$\begin{aligned} I_{m,0}(s,\kappa,\pm 1) = &\cos\left(\frac{m\pi}{2}\right) \frac{(2\pi)^m \Gamma\left(\frac{1+m}{2}\right) \Gamma\left(\frac{m-1}{2}+s\right)}{m!\Gamma(m+s)} {}_2F_1\left(\frac{m-1}{2}+s, -\frac{m}{2}+s; \frac{1}{2}; \kappa^2\right) \\ &+ m\kappa \sin\left(\frac{m\pi}{2}\right) \frac{(2\pi)^m \Gamma\left(\frac{m}{2}\right) \Gamma\left(\frac{m}{2}+s\right)}{m!\Gamma(m+s)} {}_2F_1\left(\frac{1-m}{2}, \frac{m}{2}+s; \frac{3}{2}; \kappa^2\right). \end{aligned} \tag{3.27}$$

When $m = 0$, this formula simplifies to

$$I_{0,0}(s,\kappa,\pm 1) = \sqrt{\pi} \frac{\Gamma(s-1/2)}{\Gamma(s)}. \tag{3.28}$$

The sum over c may also be evaluated exactly (see Sec. 2.5 of [22]). For $\mu = 0$ this evaluation involves the Kloosterman sum $S(0,0;c)$; from the definition (3.21), one can see that $S(0,0;c)$ is equal to the Euler totient function $\phi(c)$ (which is defined as the number of positive integers less than c that are relatively prime to c). The sum over c is a standard one

$$\sum_{c=1}^{\infty} c^{-2(m+s)} S(0,0;c) = \sum_{c>0} c^{-2(m+s)} \phi(c) = \frac{\zeta(2(m+s)-1)}{\zeta(2(m+s))}. \tag{3.29}$$

This formula can be obtained as follows. We start by noting a basic property of the totient function: for any n, $\sum_{d|n} \phi(d) = n$. In order to evaluate the sum $\sum_c c^{-\sigma}\phi(c)$, let us multiply this sum by $\zeta(\sigma) = \sum_n n^{-\sigma}$. This gives

$$\begin{aligned} \zeta(\sigma) \sum_{c=1}^{\infty} c^{-\sigma} \phi(c) &= \sum_{n,c=1}^{\infty} (nc)^{-\sigma} \phi(c) \\ &= \sum_{m=1}^{\infty} m^{-\sigma} \sum_{c|m} \phi(c) \\ &= \sum_{m=1}^{\infty} m^{1-\sigma} = \zeta(\sigma - 1). \end{aligned} \tag{3.30}$$

Setting $\sigma = 2(s+m)$ gives (3.29).

For $\mu = \pm 1$, the Kloosterman sum becomes a special case of what is known as Ramanujan's sum (see Sec. 2.5 of [22]):

$$S(\hat{n}, 0; c) = \sum_{d \in (\mathbb{Z}/c\mathbb{Z})^*} e^{2\pi i d/c} = \mu(c). \tag{3.31}$$

Here $\mu(c)$ is the Möbius function, which is defined as follows: $\mu(c) = 0$ if c is not square-free, while if c is the product of k distinct prime numbers, then $\mu(c) = (-1)^k$. The sum over c is given by

$$\sum_{c=1}^{\infty} c^{-2(s+m)} S(0, \pm 1; c) = \sum_{c=1}^{\infty} c^{-2(s+m)} \mu(c) = \frac{1}{\zeta(2(s+m))}. \tag{3.32}$$

To prove this, one may use the basic property of the Möbius function: for any n, $\sum_{d|n} \mu(d) = \delta_{n,1}$. To compute $\sum_c c^{-\sigma} \mu(c)$, we multiply this sum by $\zeta(\sigma)$ to get

$$\begin{aligned} \zeta(\sigma) \sum_{c=1}^{\infty} c^{-\sigma} \mu(c) &= \sum_{n,c=1}^{\infty} (cn)^{-\sigma} \mu(c) \\ &= \sum_{m=1}^{\infty} m^{-\sigma} \sum_{c|m} \mu(c) \\ &= \sum_{m=1}^{\infty} m^{-\sigma} \delta_{m,1} = 1. \end{aligned} \tag{3.33}$$

Setting $\sigma = 2(m+s)$ gives (3.32).

Putting these formulae together gives an exact expression for the x independent part of our Poincaré series (3.16):

$$E_0(s, \kappa, \mu) = \sum_{m=0}^{\infty} w_m(s, \kappa, \mu) y^{1-m-s}. \tag{3.34}$$

The constants $w_m(s, \kappa, \mu)$ in this expansion are independent of x and y, and are given explicitly by equations (3.26), (3.27), (3.29) and (3.32) – we will write them out in more detail for the cases of interest below. The expansion (3.34) is one of our key results. It is an explicit series expansion of our Poincaré series in powers of y^{-1}

Now we can study the behavior at $s = 1/2$ of the delicate contribution with $m = 0$. For $\mu = 0$, this term is finite because the factor of $\Gamma(s - 1/2)$ appearing in (3.26) and

the factor of $\zeta(2s)$ appearing in (3.29) both have simple poles at $s = 1/2$; as $s \to 1/2$, $\Gamma(s-1/2)/\zeta(2s) \to 2$. In fact, all of the complicated factors cancel to give

$$E_0(1/2, \kappa, 0) = -y^{1/2} + \mathcal{O}(y^{-1/2}). \tag{3.35}$$

The coefficient of the leading term is negative because $\zeta(0) = -1/2$. For $\mu = \pm 1$, the $m = 0$ term is finite because the factors of $\Gamma(s-1/2)$ in (3.28) and $\zeta(2s)$ in (3.32) both have simple poles at $s = 1/2$. We find

$$E_0(1/2, \kappa, \pm 1) = 2y^{1/2} + \mathcal{O}(y^{-1/2}). \tag{3.36}$$

Going back to (3.22), and writing simply $E(\kappa, \mu)$ for $E(1/2, \kappa, \mu)$, we now have

$$E(\kappa, 0) = y^{1/2} \exp(2\pi\kappa y) - y^{1/2} + \mathcal{O}(y^{-1/2}). \tag{3.37}$$

Actually, we have not yet discussed the Fourier modes with $\hat{n} \neq 0$; however, the integrals (3.24) vanish exponentially for $y \to \infty$ and so do not affect the assertion in eqn. (3.37). Likewise, we have

$$E(\kappa, \pm 1) = y^{1/2} \exp(2\pi(\kappa y \pm ix)) + 2y^{1/2} + \mathcal{O}(y^{-1/2}). \tag{3.38}$$

If we evaluate eqn. (3.7) for the partition function Z keeping only the first terms in eqns. (3.37) and (3.38) – the exponentially growing terms – then the formula for Z simply reduces to $Z_{0,1}$, the contribution of the Brown-Henneaux states. What we have gained from all the work that we have done is that we can now calculate corrections to $Z_{0,1}$. The leading corrections come from the corrections to the exponential terms in (3.37) and (3.38). Adding them up and evaluating (3.7), we get

$$Z = Z_{0,1} + \frac{1}{|\eta|^2}\left(-6 + \mathcal{O}(y^{-1})\right). \tag{3.39}$$

Since $1/|\eta|^2 \sim |q|^{-1/12}$, Z is governed by the Brown-Henneaux spectrum up to energy $-1/12$ (slightly below the classical black hole threshold, which is at zero energy). The number of states at that energy is not a positive integer, as one would hope, but rather -6.

Moreover, the derivation shows that the corrections are given by a power series in $1/y$, not a sum of exponentials. As was explained in Sec. 2.3, for Z to have an interpretation as Tr $\exp(-\beta H)$, the corrections must be given by exponentials.

Thus we have arrived at the main conclusion of this paper: the sum of known contributions to the partition function of pure three-dimensional gravity is not physically sensible.

It is illustrative to compute the next few terms in the expansion of the partition function (3.39) in powers of y. This means taking $m \neq 0$ in equations (3.26), (3.27), (3.29) and (3.32). Unlike the $m = 0$ case, these expressions do not have poles at $s = 1/2$, so the computation is straightforward. Let us start with the $\mu = 0$ case. We find that the coefficients $w_m(s, \kappa, \mu)$ are given by

$$w_m(1/2, \kappa, 0) = \frac{2^m \pi^{m+1/2} \zeta(2m)}{m\Gamma(m + 1/2)\zeta(2m + 1)} \kappa^m. \tag{3.40}$$

So the next few terms in the Poincaré series (3.34) are

$$E_0(1/2, \kappa, 0) = -y^{1/2} + \left(\frac{2\pi^3}{3\zeta(3)}\kappa\right) y^{-1/2} + \left(\frac{4\pi^6}{135\zeta(5)}\kappa^2\right) y^{-3/2} + \mathcal{O}(y^{-5/2}) \tag{3.41}$$

To evaluate the $\mu = \pm 1$ terms, note that at $s = 1/2$ the first two arguments of the relevant hypergeometric function appearing in (3.27) are integers. This means that the formula (3.27) simplifies considerably at $s = 1/2$ – it is just a polynomial in κ. It is

$$I_{m,0}(\frac{1}{2}, \kappa, \pm 1) = \frac{2\pi^{m+1/2}}{m\Gamma(m + 1/2)} T_m(\kappa) \tag{3.42}$$

where $T_m(\kappa)$ denotes a Chebyshev polynomial of the first kind. So the coefficients appearing in (3.34) are

$$w_m(1/2, \kappa, \pm 1/2) = \frac{2\pi^{m+1/2}}{m\Gamma(m + 1/2)\zeta(2m + 1)} T_m(\kappa). \tag{3.43}$$

This allows us to write down the next few terms in the series:

$$E_0(1/2, \kappa, \pm 1) = 2y^{1/2} + \left(\frac{4\pi}{\zeta(3)}\kappa\right) y^{-1/2} + \left(\frac{4\pi^2}{3\zeta(5)}(2\kappa^2 - 1)\right) y^{-3/2} + \mathcal{O}(y^{-5/2}). \tag{3.44}$$

Going back to (3.22), and writing $E(\kappa, \mu)$ for $E(1/2, \kappa, \mu)$, we have

$$E(\kappa, 0) = y^{1/2} \exp(2\pi\kappa y) - y^{1/2} + \left(\frac{2\pi^3}{3\zeta(3)}\kappa\right) y^{-1/2} + \left(\frac{4\pi^6}{135\zeta(5)}\kappa^2\right) y^{-3/2} + \mathcal{O}(y^{-5/2}) \tag{3.45}$$

and

$$E(\kappa, \pm 1) = y^{1/2} \exp(2\pi(\kappa y \pm ix)) + 2y^{1/2} + \left(\frac{4\pi}{\zeta(3)}\kappa\right) y^{-1/2} + \left(\frac{4\pi^2}{3\zeta(5)}(2\kappa^2 - 1)\right) y^{-3/2} + \mathcal{O}(y^{-5/2}). \tag{3.46}$$

Plugging this into (3.7) gives the expansion of the partition function

$$Z = Z_{0,1} + \frac{1}{|\eta|^2}\Big(-6 + \frac{(\pi^3 - 6\pi)(11+24k)}{9\zeta(3)} y^{-1} + \frac{5(53\pi^6 - 882\pi^2) + 528(\pi^6 - 90\pi^2)k + 576(\pi^6 - 90\pi^2)k^2}{2430\zeta(5)} y^{-2} + \mathcal{O}(y^{-3})\Big). \tag{3.47}$$

The additional contributions to the partition function in this expression have two notable features. First, and most importantly, they are not zero. Thus, as described above, the partition function truly cannot be represented as a sum of exponentials. Second, they differ qualitatively from the leading y^0 term: the additional coefficients appearing here are positive and irrational, rather than negative and integer.

3.4. $\hat{n} \neq 0$ Modes

Now we will consider the $\hat{n} \neq 0$ terms. For $\mu = 0$ and $\hat{n} \neq 0$, the integral (3.24) is a K-Bessel function

$$I_{m,\hat{n}}(s,\kappa,0) = \frac{2^{s+1}\pi^{2s+m}|\hat{n}|^{s+m-1/2}}{m!\Gamma(s+m)} y^{s+m-1/2} K_{s+m-1/2}(2\pi|\hat{n}|y). \tag{3.48}$$

The Kloosterman sum is now the general case of Ramanujan's sum

$$S(\hat{n},0;c) = \sum_{d\in(\mathbb{Z}/c\mathbb{Z})^*} e^{2\pi i \hat{n} d/c} = \sum_{\delta|\hat{n}} \mu(\delta). \tag{3.49}$$

We will simply quote the answer for the sum over c (see Sec. 2.5 of [22])

$$\sum_{c=1}^{\infty} c^{-2(s+m)} S(\hat{n},0;c) = \frac{1}{\zeta(2(s+m))} \sum_{\delta|\hat{n}} \delta^{1-2(s+m)}. \tag{3.50}$$

Taking $s = 1/2$ gives a Fourier coefficient of the regularized partition function. Consider first the $m = 0$ term. For $\hat{n} = 0$, this was the dangerous term in the analytic continuation, but for $\hat{n} \neq 0$, it simply vanishes, because the $\zeta(2s)$ in (3.50) has a pole at $c = 1/2$, causing the Kloosterman sum to vanish, and – unlike the $\hat{n} = 0$ case – the integral (3.48) is finite at $s = 1/2$. The other terms are non-zero, and give

$$E_{\hat{n}}(1/2,k,0) = \sum_{m=1}^{\infty} \frac{2^{3/2}\pi^{m+1}|\hat{n}|^m}{m!\Gamma(m+1/2)\zeta(2m+1)} \left(\sum_{\delta|\hat{n}} \delta^{-2m}\right) \sqrt{y} K_m(2\pi|\hat{n}|y). \tag{3.51}$$

Each term in the sum over m vanishes for large y as $e^{-2\pi|\hat{n}|y}$.

Unfortunately, when $\mu \neq 0$ the coefficients in the expansion of the partition function are more difficult to compute. The problem is not the integral (3.24); these integrals are finite and can be evaluated analytically, although we will not write the answer here. The answer is a finite sum of Bessel functions of the form appearing in (3.48), each of which is multiplied by a polynomial in y. As $y \to \infty$, these Bessel functions cause the integrals vanish as $e^{-2\pi|\hat{n}|y}$, just as in the $\mu = 0$ case described above.

However, when $\mu \neq 0$, the sum over c

$$\sum_{c=1}^{\infty} c^{-2(m+s)} S(\hat{n}, \mu, c), \tag{3.52}$$

though of considerable number-theoretic interest, cannot be expressed in terms of familiar number-theoretic functions such as the Riemann zeta function. Analytic properties of these sums have been extensively studied. We will simply quote the relevant results. It has been shown that the sum (3.52) defines a meromorphic function on the complex s plane. This function is essentially the Selberg zeta function associated to the modular domain $D = \mathcal{H}/SL(2,\mathbb{Z})$. When we take $s = 1/2$, the function (3.52) remains regular for elementary reasons if $m > 0$. Indeed, one can see directly that the sum converges for these values, by noting from the definition of the Kloosterman sum (3.21) that $|S(\hat{n}, \mu, c)| \leq c$. To understand the case $m = 0$, one needs deeper results that can be found in Chapter 9 of [22] (and were described to us by P. Sarnak). The key result is that the only poles of this sum, or of its analogs for other congruence subgroups of $SL(2,\mathbb{Z})$, are at the points $s + m = 1/2 + it_j$, where $t_j^2 - 1/4$ is one of the discrete eigenvalues of the hyperbolic Laplacian $\Delta = -y^2(\partial_x^2 + \partial_y^2)$ on D. A pole at $s = 1/2$, $m = 0$, will therefore arise precisely if Δ has a discrete eigenvalue at $\lambda = 1/4$. It is a happy fact that no such eigenvalue exists – the smallest discrete eigenvalue of the Laplacian on D is of order $\lambda_1 \approx 90$. Therefore, the sum (3.52) may be analytically continued to give a finite value at $s = 1/2$ for all values of m, including the dangerous case $m = 0$. In addition, simple bounds suffice to show that there is no problem with the sum over m.

This argument shows that all $\hat{n} \neq 0$ Fourier coefficients of the partition function (3.1) are finite. Together with the results of the previous subsection for $\hat{n} = 0$, this allows us to conclude that the regularization scheme described in Sec. 3.1 provides a finite answer for the partition function (3.1). We conjecture that, just as for $\hat{n} = 0$, the results are not compatible with a Hilbert space interpretation.

3.5. Aside: the Tree Level Sum over Geometries

For comparison, we will now consider the sum which arises if we neglect the one-loop contribution to the path integral described in Sec. 2.2. As we will see, in this case the answer is if anything even worse.

The sum over geometries, if we take account of only the classical action and not the one-loop correction, is

$$Z^*(\tau) = \sum_{c,d} \exp\{2\pi k \,\mathrm{Im}\,\gamma\tau\}. \tag{3.53}$$

This sum is quadratically divergent, since at large c and d the summand approaches one. Before attempting to regularize the sum, let us compare it to the "correct" sum (3.4) that does include the one-loop correction. The "correct" formula has a factor of $1/\sqrt{\mathrm{Im}\,\tau}|\eta|^2$ that is outside the summation. This factor is certainly important, but it does not affect whether the sum converges. There is also a factor of $|1-q|^2$ inside the sum. This factor also turned out in the above analysis to be less important than it may have appeared; we simply expanded it as $1-q-\bar{q}+q\bar{q}$, and wrote the partition function as a sum of four terms. All four terms were similar, and nothing particularly nice happened in adding them up.

The factor in (3.4) that actually is important, resulted from the one-loop correction, and has no analog in the "naive" sum (3.53) is the innocent-looking factor of $\sqrt{\mathrm{Im}\,\tau}|_\gamma$ that is *inside* the summation. Because of this factor, we had to evaluate our Poincaré series at $s = 1/2$.

This factor is absent in eqn. (3.53), so now, if we try to define the naive sum Z^* by introducing a parameter s as before, we will have to evaluate the resulting function at $s = 0$. In fact, the necessary s-dependent function was already introduced in eqn. (3.15); $Z^*(\tau)$ is formally given by the function $E(s,k,0)$ at $s=0$:

$$Z^*(\tau) = \lim_{s\to 0} E(s,k,0). \tag{3.54}$$

However, as an analytic function in s, $E(s,k,0)$ has a pole at $s=0$. To see this, consider the expansion (3.34) of the part of $E(s,k,0)$ which is constant in x. The $m=0$ term in this sum vanishes, because of the pole in $\Gamma(s)$ at $s=0$. The $m=1$ term gives

$$E(s,k,0) = \sqrt{\pi}\frac{\zeta(1+2s)\Gamma(s+1/2)}{\zeta(2+2s)\Gamma(1+s)} + \mathcal{O}(y^{-1}) \tag{3.55}$$

which has a pole at $s=0$ coming from the harmonic series $\zeta(1)=\infty$.

Hence, without the one-loop correction, the divergence of the sum over geometries becomes more serious. Of course, even if we could make sense of the function Z^*, we still might have trouble giving it a Hilbert space interpretation.

One might wonder what happens if we modify the definition of Z^* to include a finite subset of the Brown-Henneaux states – for example, the states of negative energy. This means that before summing over geometries, we multiply the exponential of the classical action by a polynomial $\sum_{n,m=0}^{t} a_{n,m} q^n \bar{q}^m$. The sum over geometries is a sum of terms each of which is similar to what was just described, with one such term for each monomial $q^n \bar{q}^m$. Each individual monomial will contribute a pole at $s = 0$, and generically this will survive in the sum.

4. Possible Interpretations

So far, we have analyzed the sum of known contributions to the partition function of pure three-dimensional gravity. As we learned in Sec. 3, the resulting function cannot be interpreted as Tr $\exp(-\beta H)$ for any Hilbert space operator H.

We will now address the question of how to interpret this result. The most straightforward interpretation is to take the result at face value. Three-dimensional pure gravity may not exist as a quantum theory; to get a consistent theory, it may be necessary to complete it by adding additional degrees of freedom, and there may be no canonical way to do this.

The other possibility is that some unknown contributions to the partition function should be added to the terms that we have evaluated. Here all sorts of speculations are possible. We will consider two quite different possibilities.

4.1. Cosmic Strings

Known consistent models of $2+1$-dimensional gravity with negative cosmological constant arise from string theory. For example, a famous class of models comes from Type IIB superstring theory on $\mathrm{AdS}_3 \times S^3 \times X$, where X is either a K3 surface or a four-torus.

In these models, the dimensionless ratio $k = \ell/16G$ is never a variable parameter, but always takes quantized values determined by fluxes that are chosen in the compactification. (The fact that ℓ/G is not continuously variable is actually [13] a more general consequence of the Zamolodchikov c-theorem applied to the boundary CFT.) Moreover, it

is always possible to have domain walls across which the fluxes jump. The domain walls are constructed from suitably wrapped branes.

In $2+1$ dimensions, a domain wall has a $1+1$-dimensional world-volume and so can be viewed as a cosmic string. The existence of these cosmic strings makes the models much more unified, as regions with different fluxes can appear as different domains in a single spacetime.

The usual $\mathrm{AdS}_3 \times S^3 \times X$ models have supersymmetric moduli, and the values of the string tension depend on these moduli. There is a particularly interesting supersymmetric value of the string tension at which "long strings" become possible. These are strings that can expand to an arbitrarily large size at only a finite cost of energy [24,25]. When long strings exist, the energy spectrum is continuous above a certain minimum energy, and the partition function $\mathrm{Tr}\, \exp(-\beta H)$ therefore diverges for all β.

The numerical value of the long string tension and of the excitation energy above the ground state beyond which the spectrum is continuous are proportional to the jump in $k = \ell/16G$ in crossing the string. The threshold excitation energy is of order 1 (above the ground state at energy $-2k$) if the jump in k is of order 1.

Since well-established models of three-dimensional gravity have such cosmic strings, perhaps they also present in minimal three-dimensional gravity, if it exists. In anything that one would want to call pure gravity, the string tension T measured in units of $1/\ell^2$ must go to infinity as $k \to \infty$. Otherwise, the cosmic strings would contribute excitations at the AdS_3 scale, and one would describe the model as a theory of three-dimensional gravity plus matter. We have no idea if the string tension should be proportional to k (as one might expect for solitons), to $k^{1/2}$ (as for D-branes), etc. If the jumps in k are correctly matched with T, then the strings are long strings and the partition function $\mathrm{Tr}\, \exp(-\beta H)$ that we have been trying to compute in this paper is actually divergent. If the jumps in k are smaller than this, then the partition function converges, but to compute it might involve corrections from the strings.

If one thinks that the strings should be long strings, then the requirement $T >> \ell^2$ means that the jumps in k in crossing strings are much greater than 1. There is some tension between this and the proposal in [13] that k can take any integer value. If all allowed values of k are connected by strings or domain walls (as in the known string/M-theory models), then the fact that the jumps in k are large means that the allowed values of k are sparse.

4.2. Doubled Sum Over Geometries

The scenario just described is obviously rather speculative, but at least it has the virtue of underscoring our point that the range of conceivable unknown contributions to the AdS_3 partition function is quite large.

We will now describe a quite different scenario. To motivate it, we return to the formula (2.19) for the classical action of the basic spacetime $M_{0,1}$:

$$I = -4\pi k \mathrm{Im}\,\tau. \tag{4.1}$$

We write more explicitly

$$I = 2\pi i k(\tau - \bar\tau). \tag{4.2}$$

Hence the classical approximation for the contribution of this spacetime to the partition function is $\exp(-I) = \exp(-2\pi i k(\tau - \bar\tau)) = q^{-k}\bar q^{-k}$. We notice that this is locally the product of a holomorphic function of k and an antiholomorphic function, and is globally such a product if k is an integer (ensuring that q^{-k} is single-valued). The one-loop correction preserves this factorized form, and therefore the formula (2.25) for the exact partition function $Z_{0,1}$ asociated with $M_{0,1}$ has the same properties; in fact, $Z_{0,1} = F_k(q)F_k(\bar q)$, with

$$F_k(q) = q^{-k}\prod_{n=2}^{\infty}(1-q^n)^{-1}. \tag{4.3}$$

To the extent that known formulations of three-dimensional gravity are valid, this sort of factorization holds for the contribution to the partition function of any classical geometry. See [26] for a detailed example. The gauge theory description [18,19] of three-dimensional gravity gives a natural explanation of this. With negative cosmological constant, in Lorentz signature, the gauge group is $SL(2,\mathbb{R})\times SL(2,\mathbb{R})$; the theory is a product of two decoupled $SL(2,\mathbb{R})$ theories, associated respectively with left- and right-moving modes in the boundary CFT, and this corresponds to holomorphic factorization in the Euclidean form of the theory.

In [13], it was suggested that the *exact* partition function of pure three-dimensional gravity is holomorphically factorized. It was observed that if this is the case, and the Brown-Henneaux spectrum is exact until one gets above the classical black hole threshold, then the partition function can be determined uniquely. Dual CFT's consistent with the necessary spectrum have been called extremal CFT's [27]. In subsequent work, it has been shown that the genus 2 partition function of such a CFT can be uniquely and consistently

determined [28], and there has been some work comparing it to what would be expected from three dimensions [29], but on the other hand an interesting but slightly technical argument has been given which may show that extremal CFT's do not exist [30]. Also, it has been argued [31] that extremal CFT's, if they exist, generally do not have monster symmetry for $k > 1$, in contrast to what happens [32] for $k = 1$.

Now let us discuss holomorphic factorization in view of the sum over geometries. Associated to an element

$$\gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \tag{4.4}$$

of $SL(2,\mathbb{Z})$ is a classical spacetime $M_{c,d}$. Its action is obtained by applying a modular transformation to (4.2):

$$I_\gamma(\tau) = 2\pi i k \left(\gamma\tau - \gamma\bar{\tau}\right). \tag{4.5}$$

As usual, $\gamma\tau = (a\tau + b)/(c\tau + d)$, $\gamma\bar{\tau} = (a\bar{\tau} + b)/(c\bar{\tau} + d)$. The partition function of the manifold $M_{c,d}$ is

$$Z_{c,d} = F_k(q)|_\gamma \; F_k(\bar{q})|_\gamma \,, \tag{4.6}$$

and is holomorphically factorized just like $Z_{0,1}$.

However, when we sum over geometries to evaluate the partition function

$$Z = \sum_{\gamma \in \mathcal{W}} F_k(q)|_\gamma F_k(\bar{q})|_\gamma, \tag{4.7}$$

holomorphic factorization is lost. (Here $\mathcal{W}$ is set of classical geometries $M_{c,d}$, isomorphic to the coset space $PSL(2,\mathbb{Z})/\mathbb{Z}$, with $\mathbb{Z}$ being the upper triangular subgroup of $SL(2,\mathbb{Z})$.) In fact, this formula is not simply a product of holomorphic and antiholomorphic functions, but a sum of such products, somewhat like the partition function of a rational conformal field theory. (However, in that case, each term appearing in the sum is separately invariant under $T : \tau \to \tau + 1$.)

Holomorphic factorization has been lost because the sum over topologies is a common sum for left- and right-movers. What could be added to restore holomorphic factorization? This question has a simple answer, though whether the answer is really relevant to three-dimensional gravity remains to be seen.

If we formally introduce separate topological sums for holomorphic and antiholomorphic variables, defining an extended partition function

$$\widehat{Z} = \sum_{\gamma,\gamma' \in \mathcal{W}} F_k(q)|_\gamma \; F_k(\bar{q})|_{\gamma'} \,, \tag{4.8}$$

then holomorphic factorization is restored (if the sum converges or can be regularized in a satisfactory way). After all, (4.8) can be written in the manifestly holomorphically factorized form

$$\widehat{Z} = \left(\sum_{\gamma \in \mathcal{W}} F_k(q)\big|_\gamma \right) \left(\sum_{\gamma' \in \mathcal{W}} F_k(\bar{q})\big|_{\gamma'} \right). \tag{4.9}$$

The question (apart from regularizing these sums) is to interpret the double sum over topologies.

What is the classical action corresponding to a given term in the sum in (4.9)? We can answer this question by simply applying separate modular transformations to holomorphic and antiholomorphic variables on the right hand side of eqn. (4.2). The "action" of a "classical solution" that would lead to the term in (4.8) labeled by a pair γ, γ' must be

$$I_{\gamma,\gamma'} = 2\pi i \left(\gamma\tau - \gamma'\bar{\tau}\right). \tag{4.10}$$

For $\gamma \neq \gamma'$, this formula is not real, so it is not the action of a real solution of the Einstein equations.

The most obvious way to try to interpret the formula is to interpret it as the action of a complex-valued solution of the Einstein equations. As we noted in the introduction, naively speaking the Euclidean path integral is a sum over three-manifolds M that obey the boundary conditions. But there is no established way to evaluate the contribution to the Euclidean functional integral of a given M unless M admits a classical solution of the equations of motion that one can expand around. It is because of this that our first step in Sec. 2.1 was to classify the classical solutions; what we classified were the solutions of the usual real Einstein equations with Euclidean signature.

It seems strange to simply ignore three-manifolds that do not admit classical solutions. Are their contributions meaningless or zero for some reason?

One obvious way to slightly generalize the usual framework is to consider the complexified equations of motion – the Einstein equations $R_{\mu\nu} = -\Lambda g_{\mu\nu}$, defined in the usual way except with a complex-valued but nondegenerate metric tensor $g_{\mu\nu}$. If such a complex-valued solution is given, one can hope that perturbation theory around it would make sense. In the present theory, in which perturbation theory about a classical solution is one-loop exact, one could hope to make sense of the contribution of a complex saddle point (that is, a complex-valued solution of the equations of motion) to the partition function.

Unfortunately, we have been unable to find a convincing family of solutions of the complexified Einstein equations depending on the pair $\gamma, \gamma' \in \mathcal{W}$. Since this is the case,

we think another possibility is that the nonperturbative framework of quantum gravity really involves a sum not over ordinary geometries in the usual sense, but over some more abstract structures that can be defined independently for holomorphic and antiholomorphic variables. (A similar idea is expressed in Sec. 3.1 of [29].) Only when the two structures coincide can the result be interpreted in terms of a classical geometry.

"Seeing" The Non-Classical Geometries

None of this is terribly convincing – no more so, certainly, than the discussion of cosmic strings in Sec. 4.1. However, we can describe one fact that offers some slight encouragement.

Let us examine the classical limit of the partition function, by which we mean the limit of $k \to \infty$ with fixed τ. First we will consider the ordinary partition function Z, defined by summing over ordinary geometries, and then we will consider the extended partition function $\widehat{Z}$, defined in (4.8) by a double sum whose meaning is unclear.

In the classical limit, the sum over geometries is dominated by the geometry that has the most negative classical action. In other words, we must pick γ to minimize the quantity

$$I_\gamma = 2\pi i(\gamma\tau - \gamma\bar{\tau}). \tag{4.11}$$

We can also write

$$I_\gamma = \mathrm{Re}\ (4\pi i\gamma\tau) = \mathrm{Re}\ (-4\pi i\gamma\bar{\tau})\,. \tag{4.12}$$

For generic τ, the maximum occurs for a unique γ.

There actually is a complete democracy between the different choices of γ, because they are all obtained from each other by $SL(2,\mathbb{Z})$ transformations. Each γ dominates the partition function (in the large k limit) for τ in a suitable region of the upper half plane. The thermal AdS manifold $M_{0,1}$ dominates in the usual fundamental domain $|\tau| > 1$, $|\mathrm{Re}\,\tau| < 1/2$. Across the arc $|\tau| = 1$, $\mathrm{Re}\,\tau \leq 1/2$, there is a Hawking-Page phase transition to a region dominated by $M_{1,0}$, which is the Euclidean black hole. The upper half plane is tesselated in phases dominated by different classical solutions. Details are further discussed in Sec. 6.

Now let us consider the extended partition function $\widehat{Z}$. The action, given in eqn. (4.10) now depends on the pair γ, γ', and is complex. What we want to minimize is the real part of the action:

$$\mathrm{Re}\, I_{\gamma,\gamma'} = \mathrm{Re}\ (2\pi i(\gamma\tau - \gamma'\bar{\tau})) = \mathrm{Re}\ (2\pi i\gamma\tau) + \mathrm{Re}\ (-2\pi i\gamma'\bar{\tau}) = \mathrm{Re}\ (2\pi i\gamma\tau) + \mathrm{Re}\ (2\pi i\gamma'\tau)\,. \tag{4.13}$$

The minimum is always at $\gamma = \gamma'$. In fact, we can use eqn. (4.12) to show that

$$\mathrm{Re}\, I_{\gamma,\gamma'} = \frac{1}{2}\left(I_\gamma(\tau) + I_{\gamma'}(\tau)\right). \tag{4.14}$$

Whatever is the minimum with respect to x of $I_x(\tau)$, the minimum of $\mathrm{Re}\, I_{\gamma,\gamma'}(\tau)$ is at $\gamma = \gamma' = x$.

The conclusion is that, even if non-classical geometries exist and the correct partition function is a sum over non-classical geometries as well as classical ones, the semiclassical limit of large k is always dominated by a classical geometry. Non-classical geometries exist, but they never dominate the semiclassical limit.

What do we have to do to "see" a non-classical geometry in the semiclassical limit? Naturally, we have to ask a non-classical question. The partition function as we have defined it so far is $Z(\tau) = \mathrm{Tr}\, \exp(2\pi i \tau L_0 - 2\pi i \bar\tau \widetilde{L}_0)$, where τ is a point in the upper half plane $\mathcal{H}$ and $\bar\tau$ is its complex conjugate. In terms of $\widetilde\tau = -\bar\tau$, which also takes values in $\mathcal{H}$, we have $Z(\tau) = \mathrm{Tr}\, \exp(2\pi i \tau L_0 + 2\pi i \widetilde\tau \widetilde{L}_0)$. Now instead of defining $\widetilde\tau$ as $-\bar\tau$, let us relax this condition and simply think of $\widetilde\tau$ as a second point in $\mathcal{H}$.[5] Then we should rewrite (4.14) in the form

$$\mathrm{Re}\, I_{\gamma,\gamma'} = \frac{1}{2}\left(I_\gamma(\tau) + I_{\gamma'}(\widetilde\tau)\right). \tag{4.15}$$

Now it is clear that given any pair γ, γ', we can pick the pair $\tau, \widetilde\tau$ so that the extended partition function $\widehat{Z}$ is dominated in the semiclassical limit by γ, γ'. We simply pick τ so that $I_x(\tau)$ is minimized for $x = \gamma$, and $\widetilde\tau$ so that $I_x(\widetilde\tau)$ is minimized for $x = \gamma'$.

Thus, for each non-classical geometry defined by a pair γ, γ', we can find a question that that geometry dominates in the semiclassical limit. But to do this, we have to ask a rather exotic question that itself depends on a rather unusual analytic continuation.

5. Black Hole Entropy and its Corrections

In this and the following section we will discuss a few implications of the assumption of holomorphic factorization. In this section we will discuss black hole entropy, and how it can be reproduced in holomorphically factorized theories.

[5] The partition function of a theory in which the high energy density of states is given by the entropy of a BTZ black hole remains convergent after this analytic continuation to general $\tau, \widetilde\tau \in \mathcal{H}$.

In Sec. 2, we computed the perturbative corrections to the saddle point action for the geometry $M_{0,1}$. Using a modular transformation, this leads to a new formula for the subleading corrections to the entropy of the BTZ black hole. We will compare these new subleading corrections with those that occur in the extremal holomorphic partition functions of [13].

5.1. Subleading Corrections to the Bekenstein-Hawking Formula

As we noted in Sec. 2, the geometry we called $M_{1,0}$ is just the Euclidean continuation of the BTZ black hole. The modular parameter τ is related to the Hawking temperature β^{-1} and angular potential θ of the black hole, by $\tau = \theta + i\beta$. The black hole partition function is found by applying the modular transformation $\tau \to -1/\tau$ to the expression (2.23) for the partition function $Z_{0,1}$ of thermal AdS. This gives the one-loop corrected partition function for the Euclidean BTZ black hole

$$Z_{1,0} = \mathcal{Z}(\tau)\bar{\mathcal{Z}}(\bar{\tau}) \tag{5.1}$$

where we have defined the holomorphic piece of the BTZ partition function

$$\mathcal{Z}(\tau) = \frac{q_-^{-(k-1/24)}(1-q_-)}{\eta(-1/\tau)}. \tag{5.2}$$

In this formula we have defined $q_- = e^{-2\pi i/\tau}$. It will also be useful to write $\mathcal{Z}(\tau)$ as

$$\mathcal{Z}(\tau) = \sum_{\Delta=-k}^{\infty} C_\Delta q_-^\Delta \tag{5.3}$$

where the coefficients C_Δ are

$$C_\Delta = p(\Delta' + k) - p(\Delta' + k - 1). \tag{5.4}$$

Here $p(N)$ denotes the number of partitions of the integer N.

Equation (5.2) encodes various quantum corrections to the thermodynamic properties of the BTZ black hole. Equation (5.2) is a canonical ensemble partition function, so one may, for example, compute the black hole entropy using the usual formula

$$S(\beta,\theta) = \log\, Z_{1,0} - \beta\; Z_{1,0}^{-1}\; \frac{\partial Z_{1,0}}{\partial \beta} \tag{5.5}$$

where $\tau = \theta + i\beta$. This is a canonical ensemble entropy, evaluated at fixed temperature β^{-1} and angular potential θ.

For the purpose of comparing with CFT predictions, however, it is more useful to compute the microcanonical entropy, which counts the number of states $N(M, J)$ at fixed energy M and angular momentum J. The energy M and angular momentum J of a state are related to left and right-moving dimensions of the state by[6]

$$\begin{aligned} M &= \Delta + \bar{\Delta} \\ J &= \Delta - \bar{\Delta}. \end{aligned} \tag{5.6}$$

So we may write this density of states as $N(\Delta, \bar{\Delta})$. This density of states is computed from the partition function $Z_{1,0}(\tau, \bar{\tau})$ by a pair of Laplace transforms

$$N(\Delta, \bar{\Delta}) = \left(\int_{i\epsilon - \infty}^{i\epsilon + \infty} d\tau \right) \left(\int_{i\epsilon - \infty}^{i\epsilon + \infty} d\bar{\tau} \right) q^{-\Delta} \bar{q}^{-\bar{\Delta}} Z_{1,0}(\tau, \bar{\tau}). \tag{5.7}$$

These are the usual Laplace transforms that appear when going from canonical to microcanonical ensemble. We should emphasize that in this expression τ and $\bar{\tau}$ should be regarded as independent variables.

We should note that, since (5.2) is a semiclassical partition function around a black hole background, it will not have an exact interpretation as a quantum mechanical trace. Moreover, the quantization of Δ and $\bar{\Delta}$ will not be visible in this approximation.[7] So equation (5.7) should be thought of only as a semi-classical approximation to the number of states with dimension $(\Delta, \bar{\Delta})$ in the exact quantum theory. Still, we want to explore this semiclassical formula.

The microcanonical entropy $S(\Delta, \bar{\Delta})$ is just

$$S(\Delta, \bar{\Delta}) = \log\, N(\Delta, \bar{\Delta}). \tag{5.8}$$

Of course, since $Z_{1,0}$ is holomorphicaly factorized, we may write

$$N(\Delta, \bar{\Delta}) = \mathcal{N}(\Delta)\mathcal{N}(\bar{\Delta}), \qquad S(\Delta, \bar{\Delta}) = \mathcal{S}(\Delta) + \mathcal{S}(\bar{\Delta}) \tag{5.9}$$

[6] We are working in AdS units, with $\ell = 1$.

[7] To be more precise, in an exact theory with finite entropy, $N(\Delta, \bar{\Delta})$ will be a sum of delta functions of $\Delta, \bar{\Delta}$ multiplying positive integers; but the semiclassical approximation does not have this form.

where

$$\mathcal{N}(\Delta) = \int_{i\epsilon-\infty}^{i\epsilon+\infty} d\tau\, q^{-\Delta} \mathcal{Z}(\tau). \tag{5.10}$$

We may use the expansion (5.3) to get

$$\begin{aligned} \mathcal{N}(\Delta) &= \sum_{\Delta'=-k}^{\infty} C_\Delta \int_{i\epsilon-\infty}^{i\epsilon+\infty} d\tau q^{-\Delta} q_-^{\Delta'} \\ &= \sum_{\Delta'=-k}^{\infty} C_\Delta \int_{i\epsilon-\infty}^{i\epsilon+\infty} d\tau \exp\left\{-2\pi i\left(\Delta\tau + \frac{\Delta'}{\tau}\right)\right\}. \end{aligned} \tag{5.11}$$

To do this integral, recall the following representation of the Bessel function as a contour integral:

$$I_1(z) = \frac{1}{2\pi i} \oint t^{-2} e^{(z/2)(t+t^{-1})} dt \tag{5.12}$$

where the contour encloses the origin in a counterclockwise direction. Taking $t \to 1/t$ and letting $z = 4\pi\sqrt{-\Delta\Delta'}$, this becomes the integral appearing in (5.11). We end up with the following formula for the microcanonical entropy:

$$\mathcal{N}(\Delta) = e^{\mathcal{S}(\Delta)} = 2\pi \sum_{\Delta'=-k}^{\infty} C_\Delta \sqrt{\frac{-\Delta'}{\Delta}} I_1(4\pi\sqrt{-\Delta\Delta'}) \tag{5.13}$$

where C_Δ was defined in (5.4). Formula (5.13) is the main result of this section.

In the semiclassical approximation, this formula will be dominated by the term with $\Delta' = -k$. We may then use the asymptotic formula for the Bessel function

$$I_1(z) = \frac{1}{\sqrt{2\pi z}} e^z \left(1 - \frac{3}{8} z^{-1} + ...\right), \qquad \text{at } z \to \infty \tag{5.14}$$

to get

$$\mathcal{S}(\Delta) = \log \mathcal{N}(\Delta) = 4\pi\sqrt{k\Delta} + \frac{1}{4}\log\, k - \frac{3}{4}\log\, \Delta - \frac{1}{2}\log\, 2 + \ldots \tag{5.15}$$

The first term is the usual Bekenstein-Hawking term, proportional to the area of the BTZ black hole. The other terms are logarithmic corrections that typically appear when the entropy is computed in microcanonical, as opposed to a canonical, ensemble. They were computed for the BTZ black hole in [33].

Of course, equation (5.13) contains many interesting subleading terms in addition to these logarithmic terms. Let us consider the terms in (5.13) with $\Delta > -k$. For large values of Δ/k, which is to say for black holes whose radius is large in AdS units, these terms are

exponentially subleading. To see this, let us compare the $\Delta = -k$ and $\Delta = -k+1$ terms in the sum using the asymptotic behavior of the Bessel function (5.14). We find

$$\frac{I_1(4\pi\sqrt{k\Delta})}{I_1(4\pi\sqrt{(k-1)\Delta})} \sim e^{2\pi\sqrt{\Delta/k}}, \qquad \text{as } \Delta k \to \infty. \tag{5.16}$$

Thus for Δ/k large the sum is dominated by the $\Delta = -k$ term plus terms which are exponentially small in $\sqrt{\Delta/k}$. For black holes whose size is of order the AdS scale, these additional terms can become relevant.

We should make one more important comment about the structure of the subleading terms in (5.13). The terms with $\Delta' < 0$ are qualitatively of the same form as the leading contribution described above. The terms with $\Delta' > 0$ are qualitatively different: for these terms the factors of $\sqrt{-\Delta'}$ appearing in (5.13) are imaginary, and we must worry, for example, about which branch of the solution (5.13) we should take. At this point one might wonder about the physical interpretation and implications of these funny terms. However, as we will see below, it is precisely these unusual-looking terms which are absent in the correct microscopic computation of the black hole entropy. So we will not worry too much about their appearance here.

Our goal is to understand to what extent the microcanonical entropy formula (5.13) can be reproduced in a holomorphically factorized partition function. Of course, we could attempt to reproduce the canonical formula (5.5) instead, as these two expressions contain precisely the same information. But we will find the problem of computing (5.13) to be technically simpler. Our conclusion will be that the terms in (5.13) with $-k \leq \Delta' < 0$ are reproduced in the full partition function.

Before proceeding, let us ask under what circumstances we expect the formula (5.13) for the entropy to be a good approximation to the number of states of the exact quantum theory. Of course, we must take k to be large, in order for the semi-classical approximation to be valid. In addition, we must take Δ to be large so that the corresponding black hole is large in Planck units and dominates the entropy of the system.

5.2. The Rademacher Expansion for Holomorphic Partition Functions

Let us now consider a modular-invariant, holomorphic CFT with central charge $c = 24k$. For such CFTs, k must be an integer, so the holomorphic part of the partition function can be expanded as

$$Z(\tau) = \sum_{\Delta=-k}^{\infty} F_\Delta q^\Delta \tag{5.17}$$

where the F_Δ are some positive integers. Modular invariance fixes $Z(\tau)$ in terms of the finite set of coefficients $\{F_\Delta | -k \leq \Delta \leq 0\}$. Once these polar coefficients are fixed, an explicit formula can then be given for $Z(\tau)$ as a polynomial in the j-invariant, although we will not write this formula here. The important point is that all of the coefficients F_Δ with $\Delta > 0$ are fixed in terms of the coefficients F_Δ with $\Delta \leq 0$. The formula for the F_Δ with $\Delta > 0$ in terms of these polar terms is

$$F_\Delta = 2\pi \sum_{-k \leq \Delta' < 0} \sqrt{\frac{|\Delta'|}{\Delta}} F_{\Delta'} \sum_{n=1}^{\infty} \frac{1}{n} S(\Delta, |\Delta'|, n) I_1\left(\frac{4\pi}{n}\sqrt{\Delta|\Delta'|}\right) \tag{5.18}$$

where $S(\Delta, \Delta', n)$ is the Kloosterman sum defined in (3.21). This is a convergent series expansion of F_Δ. Expansions of the form (5.18) are known as Rademacher expansions. Rademacher expansions were first applied to three-dimensional gravity in [10].

In [13], a conjecture was made regarding the form of the partition function (5.17). The conjecture is that the theory contains no primary fields with $-k < \Delta \leq 0$. CFTs with this property are known as extremal CFTs, since in a sense this is the most extreme conjecture possible: modular invariance forces one to include additional primaries with $\Delta > 0$. In terms of the coefficients F_Δ, the conjecture of [13] is that

$$F_\Delta = C_\Delta \qquad \text{for } -k < \Delta \leq 0. \tag{5.19}$$

Here C_Δ are the perturbative coefficients defined in (5.4).

The extremal partition function (or any similar function in which the number of primary fields of dimension $\leq k$ is not too large) is consistent with the leading Bekenstein-Hawking term in the black hole entropy for Δ, k large. We are now in a position to ask whether subleading terms given by (5.13) are reproduced as well. First, we note that the sums (5.18) and (5.13) are quite similar. In particular, when $n = 1$ the Kloosterman sum is 1. In this case the summand in (5.18) coincides precisely with that in (5.13), at least when $\Delta' < 0$. Thus the Rademacher expansion (5.18) contains the first k terms in the one-loop corrected black hole entropy (5.13). This is the main result of this section.

There are two important differences between (5.18) and (5.13). The first is that the sum in (5.13) is over $-k < \Delta' < \infty$, while the sum (5.18) is only over $-k < \Delta' < 0$. So the funny terms in (5.13) with $\Delta' > 0$ are absent in the full microscopic entropy. It would be interesting to understand the physical interpretation of these terms, but we will not attempt to do so here.

The second important difference is that the Rademacher expansion (5.18) includes terms with $n > 1$. These terms provide exponentially small corrections to the black hole entropy. To see this, we may again use the asymptotic behavior of the Bessel function (5.14). For black holes with large Δ the ratio of the $n = 1$ and $n = 2$ terms in the sum (5.18) goes like

$$\frac{I_1(2\pi\sqrt{\Delta|\Delta'|})}{I_1(4\pi\sqrt{\Delta|\Delta'|})} \sim e^{-2\pi\sqrt{\Delta|\Delta'|}}. \tag{5.20}$$

So at large Δ, the $n > 1$ terms in the Rademacher expansion are exponentially small in Δ. Again, it would be interesting to study the physical interpretation of these terms.

6. The Hawking-Page Phase Transition

In this section we address an apparent puzzle posed by holomorphic factorization and the Hawking-Page transition [15]. As we will see, the resolution to this puzzle is that if holomorphic factorization is valid, then the Hawking-Page phase transition is described by a condensation of Lee-Yang zeroes of the partition function [34,35]. We examined the zeroes of the partition function as a result of a question from M. Kaneko, and the zeroes we find are similar to what have been obtained in investigations [36,37] of certain somewhat similar modular functions.

The partition function $Z(\tau)$ computes a canonical ensemble partition function at fixed temperature Im τ and angular potential Re τ. As described in Sec. 2, for each value of τ, an infinite family of classical geometries $M_{c,d}$ will contribute to $Z(\tau)$. The pair of integers (c, d) labels an element of the coset $SL(2,\mathbb{Z})/\mathbb{Z}$, where $\mathbb{Z}$ denotes the shifts $\tau \to \tau + n$. In Sec. 2.2, we computed the contribution of $M_{c,d}$ to the partition sum $Z(\tau)$. If we choose an element $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$ of $SL(2,\mathbb{Z})$ corresponding to $M_{c,d}$, the contribution to the partition function was given by $Z_{0,1}(\frac{a\tau+b}{c\tau+d})$.

For a given value of τ, one can ask which geometry $M_{c,d}$ has the largest contribution to $Z(\tau)$ in the semiclassical limit $k \to \infty$. As explained in [9,10] and in Sec. 4.2, this question can be answered by maximizing the classical action $-4\pi \mathrm{Im}\,(\gamma\tau)$ as a function of γ. If τ is in the usual fundamental domain $|\tau| > 1$, $|\mathrm{Re}\,\tau| \leq 1/2$, or any of its translates by $\tau \to \tau + n$, then the answer is that the dominant classical solution is $M_{0,1}$, which describes thermal AdS space. For any τ, the dominant classical solution can be found by asking which value of $\gamma \in SL(2,\mathbb{Z})$ will take τ into the fundamental domain or one of its translates under $\tau \to \tau + n$. For example, $M_{1,0}$, whose Lorentzian continuation describes a black

hole in equilibrium with its Hawking radiation, dominates whenever there exists an n such that $-1/(\tau+n)$ lies in the fundamental domain. Roughly speaking, this is a regime of low temperature and angular potential. More complicated geometries $M_{c,d}$, which do not have simple Lorentzian interpretations (as they do not admit an analytic continuation to a real solution with Lorentz signature) dominate in other regions of the upper half plane $\mathcal{H}$. Putting this together, we find that the phase diagram of three-dimensional gravity, in the semiclassical limit, is given by a sort of subtesselation of the usual tesselation of $\mathcal{H}$ by fundamental domains of $SL(2,\mathbb{Z})$.

We get a subtesselation because certain boundaries between fundamental domains can be crossed without any jump in the dominant geometry (an example is that, for $\mathrm{Im}\,\tau >> 1$, there is no jump in crossing between domains related by $\tau \to \tau+n$). The usual tesselation of $\mathcal{H}$ by fundamental domains is sketched in Fig. 3a, and the subtesselation relevant to three-dimensional gravity is sketched in Fig. 3b.

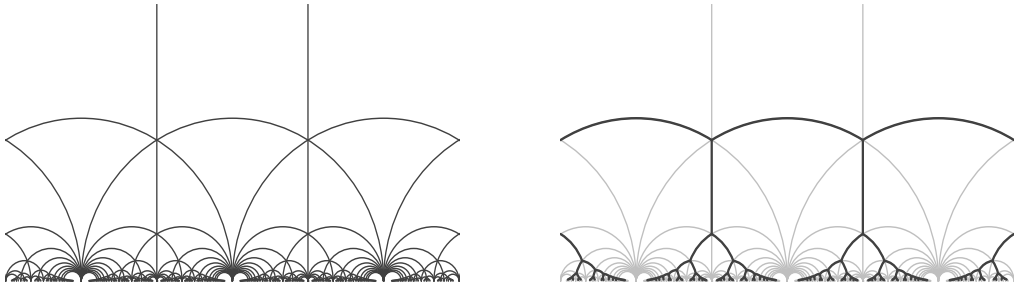


Figure 3: a) The standard tesselation of the upper half τ-plane by $SL(2,\mathbb{Z})$ fundamental regions. b) The subtesselation that represents the phase diagram of three dimensional gravity. The phase boundaries, represented by solid black arcs, connect fixed points of $SL(2,\mathbb{Z})$ of order 3. This characterization was found by considering special cases and then using modular invariance.

It is believed that the phase structure as a function of τ of a wide range of three-dimensional gravity theories (with different sets of fields) is as we have just described. For any finite value of the dimensionless ratio $k = \ell/16G$, the partition function is smooth as a function of τ, but for $k \to \infty$ or $G \to 0$, it becomes non-smooth along the curves drawn in Fig. 3b.

We will address here a puzzle that arises if one assumes holomorphic factorization. In this case, we take k to be a positive integer. The partition function factorizes as

$\widehat{Z} = Z_k(q)Z_k(\bar{q})$, where the function Z_k is a holomorphic and modular-invariant function. Z_k has an expansion near $q = 0$ involving the negative energy Brown-Henneaux states,

$$Z_k = q^{-k} + \dots, \tag{6.1}$$

and has no other singularities. In order for $\widehat{Z}$ to exhibit the phase structure sketched in Fig. 3b in the semiclassical limit, Z_k must do the same.

The poses a puzzle, since Z_k is holomorphic. There is no problem for a sequence of smooth but not holomorphic functions depending on a parameter to become non-smooth along a phase boundary in some limit. But how can this happen to a sequence of holomorphic functions?

This question was actually answered by Lee and Yang [34,35]. The original idea of Lee and Yang is that although a system in finite volume can have no phase transition, its partition function, in its dependence on the complexified thermodynamic variables, can have zeroes. Then, in the infinite volume limit, the zeroes become more numerous and may become dense or "condense" along a certain arc, and a true phase transition can emerge.

In our problem, the limit $k \to \infty$ is analogous to a thermodynamic limit, since the effective size of AdS space (accessible at a given temperature) grows with k. In a fundamental domain compactified by adding its "cusp," the function Z_k has precisely k poles, exhibited in (6.1). Hence the number of zeroes in a fundamental domain is also equal to k. Here in the case of zeroes that are on the boundary of the fundamental domain, one must count only half the zeroes (or one half of the boundary) to avoid double-counting.

Before going into details, we will give an informal explanation in which we ignore the one-loop correction. We consider the phase boundary along the arc $|\tau| = 1$, $|\mathrm{Re}\,\tau| \leq 1/2$ separating the thermal AdS phase from the Euclidean black hole. We will denote the portion of this arc with $\mathrm{Re}\ \tau \geq 0$ by C:

$$C = \{\tau = x + iy \big| |\tau| = 1, 0 \leq x \leq 1/2\}. \tag{6.2}$$

The other portion of the arc, with $\mathrm{Re}\ \tau \leq 0$, is the image of C under the modular transformation $\tau \to -1/\tau$. Now, the contributions of $M_{0,1}$ and $M_{1,0}$ to the holomorphic Poincaré series (ignoring the one-loop correction) are $\exp(-2\pi ik\tau)$ and $\exp(-2\pi ik(-1/\tau))$. On the arc C, we have $1/\tau = \bar{\tau}$ and the sum of these two contributions is

$$\exp(-2\pi ik\tau) + \exp(2\pi ik\bar{\tau}). \tag{6.3}$$

The sum vanishes if $\exp(2\pi ik(\tau+\bar\tau)) = -1$. Setting $\tau = \exp(i\theta)$, $\pi/3 \le \theta \le 2\pi/3$, we need $\exp(4\pi ik\cos\theta) = -1$ or $4\pi k\cos\theta = (2n+1)\pi$, $n \in \mathbb{Z}$. On the arc C, we have $0 \le \cos\theta \le 1/2$. So this equation has k roots. In the limit $k \to \infty$, these roots become dense everywhere along the arc. This is the condensation of Lee-Yang zeroes of the partition function.

Now we give a more precise account including the one-loop correction. However, the result does not depend on all the details of the function Z_k. We will show, following [36,37], that for a certain class of holomorphic functions, all zeroes lie on the arc C or one of its images under $SL(2,\mathbb{Z})$, as shown in Fig. 3. In the large k limit, the zeroes become dense on C.

We consider a modular-invariant function Z_k that is regular except for a pole of order k at $q = 0$. Such a function has a Laurent expansion

$$Z_k(\tau) = \sum_{\Delta=-k}^{\infty} F_\Delta q^\Delta \,. \tag{6.4}$$

We require that $F_{-k} = 1$ and that the coefficients F_Δ, $\Delta \le 0$, do not increase too quickly. More precisely, we will show that if

$$F_\Delta < e^{2\pi(0.61)(\Delta+k)}, \qquad \text{for } \Delta \le 0 \tag{6.5}$$

then the zeroes are all as just described. The origin of the numerical coefficient will be described below. This bound has a simple physical explanation: when there are too many primary fields of low dimension, i.e. too much light matter in the theory, the formation of stable black holes is not possible. Problems with black hole physics when there are many light species of matter have been considered in, for example, [38].

We should emphasize that the extremal partition functions proposed in [13] obey the bound (6.5). So in the large k limit the extremal partition functions exhibit the condensation of zeroes described above. In fact, for $\Delta < 0$ and large k the coefficients of the extremal partition function grow much more slowly than (6.4). For large $\Delta + k$, they grow like

$$F_\Delta \sim e^{\frac{\pi}{6}\sqrt{\Delta+k}} << e^{2\pi(0.61)(\Delta+k)}. \tag{6.6}$$

So one could add many more additional primaries with dimension $\Delta < 0$ without spoiling the phase transition described above.

Our proof of the foregoing assertions relies on a few specific facts about modular functions, which we now review.

6.1. Review: Properties of J

First, we recall a few properties of the modular invariant $J(\tau)$. Proofs of many of our assertions in this and the next section can be found in [39]. The J function has a $q-$expansion

$$J(\tau) = \sum_{m \geq -1} c(m) q^m = \frac{1}{q} + 196884q + \dots \tag{6.7}$$

where the coefficients $c(m)$ are positive integers. Our choices of overall normalization and constant term $c(0) = 0$ agree with [32] but differ from the most classical convention. As the $c(m)$ are positive, it follows that J obeys

$$J(\tau) = \bar{J}(-\bar{\tau}) \tag{6.8}$$

and so is real along the imaginary τ axis. In particular, it can be shown that J decreases monotonically along the imaginary axis from $J(i\infty) = \infty$ to a minimum at $\tau = i$ with $J(i) = 984$. (Continuing past i on the imaginary axis, J grows again in view of its invariance under $\tau \to -1/\tau$.)

The modular domain

$$D = \{\tau = x + iy \big| |\tau| \geq 1, -1/2 \leq x \leq 1/2\} \tag{6.9}$$

is a fundamental region for $SL(2,\mathbb{Z})$. On the interior of D, J takes every complex value exactly once. (This follows from the fact that in the compactified fundamental domain, J has precisely one pole, which is at $\tau = i\infty$.) Using the invariance of J under $SL(2,\mathbb{Z})$ and (6.8), one can show that $J(\tau)$ takes real values on the boundary ∂D. In particular, it turns out that on the arc $\tau = e^{i\theta}$, $J(\tau)$ decreases monotonically from 984 to -744 as θ runs from $\pi/2$ to $\pi/3$. This is the arc we called C above. On the arcs $x = \pm 1/2$, J decreases monotonically from -744 to $-\infty$ as y runs from $\sqrt{3}/2$ to ∞.

Although the exact form of $J(q)$ is quite complicated, in many cases the "tree-level approximation"

$$J \sim q^{-1} + \dots \tag{6.10}$$

is very useful. To this end, we note that for any given value of $\tau = x + iy$, we have the bound

$$|J(\tau) - q^{-1}| = |\sum_{m \geq 1} c(m) q^m| \leq \sum_{m \geq 1} c(m) |q|^m = J(iy) - e^{2\pi y}. \tag{6.11}$$

The function $\sum_{m\geq 1} c(m)|q|^m$ depends only on y and is a monotonically decreasing function of y, so in the fundamental domain D, it is bounded above by its value at $y = \sqrt{3}/2$, which is the minimum of y in D. This value is

$$M = J(i\sqrt{3}/2) - e^{\pi\sqrt{3}} \approx 1335. \tag{6.12}$$

So in D, we have

$$|J(\tau) - q^{-1}| \leq M. \tag{6.13}$$

Applying the triangle inequality, it follows that, throughout D, we can bound the value of J by

$$|J(\tau)| \leq e^{2\pi y} + M. \tag{6.14}$$

The approximation (6.10) is a good one when the imaginary part of τ is large, i.e. when τ is close to the cusp at $\tau \to \infty$. By considering modular transformations of the equation (6.13) we arrive at other approximations which are good in other regions of the upper half plane. In particular, for any element $\gamma \in SL(2,\mathbb{Z})$, if $\gamma\tau$ is in the fundamental domain D we have

$$|J(\tau) - e^{-2\pi i\gamma\tau}| \leq M, \tag{6.15}$$

and therefore

$$|J(\tau)| \leq e^{2\pi \mathrm{Im}\ \gamma\tau} + M. \tag{6.16}$$

This final bound is particularly useful, for the following reason. For a given value of τ, consider the set of possible values of Im $\gamma\tau$, for all $\gamma \in SL(2,\mathbb{Z})$. This set has a finite maximum value which occurs when $\gamma\tau$ is in the fundamental domain or one of its images under the map $\tau \to \tau + n$. So, using the formula (3.9) for Im $\gamma\tau$, we conclude that for any value of $\tau = x + iy$ in the upper half plane we have the bound

$$|J(\tau)| \leq \exp\left\{\max_{c,d} \frac{2\pi\,\mathrm{Im}\ \tau}{|c\tau + d|^2}\right\} + M. \tag{6.17}$$

Here the maximum is taken over all relatively prime c and d.

6.2. Properties of $T_n J$

We will now consider the action of the Hecke operator T_n on J, defined by

$$T_n J = \sum_{\delta|n} \sum_{\beta=0}^{\delta-1} J((n\tau + \beta\delta)/\delta^2). \tag{6.18}$$

This is a new modular invariant function with an n-th order pole at $q = \infty$. Our normalization (which differs by a factor of n from that in the literature) has been chosen so that the $q-$expansion of $T_n J$ starts with q^{-n}:

$$T_n J = \sum_{m=-n}^{\infty} c_n(m) q^m = q^{-n} + \mathcal{O}(q). \tag{6.19}$$

The Fourier coefficients $c_n(m)$ of $T_n J$ are related to the coefficients $c(m)$ of J by

$$c_n(m) = \sum_{\delta|(m,n)} \frac{n}{\delta} c\left(mn/\delta^2\right) q^m. \tag{6.20}$$

From this it follows that the Fourier coefficients of $T_n J$ are positive integers, just like those of J.

From these facts, it is straightforward to see that $T_n J$ satisfies many of the same properties as J. In particular, $T_n J$ is real along the imaginary axis, as well as on the boundary of the fundamental domain ∂D. Moreover, just as with J, the "tree-level approximation"

$$T_n J \sim q^{-n} + \ldots \tag{6.21}$$

is good enough for many purposes. This approximation is very good when $\tau \to i\infty$, but actually it can be extended to understand the zeroes of $T_n J$, as was shown in [37]. We will summarize their argument here. In section 6.3, we apply the results derived below to more general modular-invariant partition functions.

The basic idea is to apply either (6.15) or (6.17) to each term in the sum (6.18). This allows us to bound the value of $T_n J(\sigma)$ for a point $\sigma = x + iy$ on the arc C defined by $|\sigma| = 1$ and $0 \le x \le 1/2$. Let us first consider the $\delta = 1$ term in (6.18), which is $J(n\sigma)$. For any σ in C, we can find an integer m such that $n\sigma + m$ lies in the fundamental domain. This allows us to apply (6.15) with $\tau = n\sigma$ and $\gamma\tau = n\sigma + m$ to get

$$|J(n\sigma) - e^{-2\pi i n\sigma}| \le M. \tag{6.22}$$

Now, consider the $\delta = n$, $\beta = 0$ term in (6.18), which is $J(\sigma/n)$. Since σ lies in C, we can find an integer m such that $-n/\sigma + m$ lies in the fundamental domain. So we may apply (6.15), with $\tau = \sigma/n$ and $\gamma = -n/\sigma + m$ to get

$$|J(\sigma/n) - e^{-2\pi i n/\sigma}| \leq M. \tag{6.23}$$

Using these two equations, we may apply the triangle inequality to (6.18) to get

$$|T_n J(\sigma) - e^{-2\pi i n\sigma} - e^{-2\pi i n/\sigma}| \leq 2M + \sum{}' \left| J\left(\frac{n\sigma + \beta\delta}{\delta^2}\right)\right|. \tag{6.24}$$

Here the sum $\sum'$ is over the same set of β and δ as in (6.18), except we have dropped the two terms with $\delta = 1$ and $\delta = n, \beta = 0$. We will now apply (6.17) to each term on the right hand side of (6.24). Consider first the term with $\delta = n$, $\beta = n-1$, which is $J\left(\frac{\sigma-1}{n}\right)$. We would like to apply (6.17) with $\tau = \frac{\sigma-1}{n}$, so we must ask what possible values $|c\tau + d|$ can take for this value of τ. Now, since σ lies on C, it follows that $|\sigma - 1| > 1$ and hence $|\tau| = |\frac{\sigma-1}{n}| > 1/n$. This implies that $|c\tau + d| \geq 1/n$ for all possible choices of c and d. Since Im $\tau = y/n$, equation (6.17) gives

$$\left| J\left(\frac{\sigma + n - 1}{n}\right)\right| \leq e^{2\pi n y} + M. \tag{6.25}$$

Let us consider the case where $\delta = n$ and $0 < \beta < n-1$, where we may apply (6.17) with $\tau = \frac{\sigma+\beta}{n}$. For this range of β, $|\sigma + \beta| > \sqrt{2}$, so $|\tau| > \sqrt{2}/n$. Hence $|c\tau + d| > \sqrt{2}/n$ and (6.15) gives

$$\left| J\left(\frac{\sigma + \beta}{n}\right)\right| \leq e^{\pi n y} + M. \tag{6.26}$$

Finally, we consider the cases where $1 < \delta < n$. In this case we apply (6.17) with $\tau = (n\sigma + \beta\delta)/\delta^2$. The fact that σ lies on C implies that $|c\tau + d| > \sqrt{3}n/\delta^2$. So we end up with the bound

$$\left| J\left(\frac{\sigma + \beta}{n}\right)\right| \leq e^{\frac{2\pi n}{3} y} + M. \tag{6.27}$$

Putting this all together, equation (6.24) becomes

$$|T_n J(\sigma) - e^{-2\pi i n\sigma} - e^{-2\pi i n/\sigma}| \leq e^{2\pi n y} + n e^{\pi n y} + n^2 e^{\frac{2\pi n}{3} y} + n^2 M. \tag{6.28}$$

The factors of n and n^2 in (6.28) come from the simple fact that there are less than n terms with $\delta = n$, and less than n^2 terms with $1 < \delta < n$. Multiplying both sides of (6.28) by $e^{-2\pi n y}$, and using the fact that $e^{-2\pi i n/\sigma} = e^{2\pi i n\bar{\sigma}}$ for points on the curve C, this becomes

$$|T_n J e^{-2\pi n y} - 2\cos(2\pi n x)| \leq 1 + n e^{-\pi n y} + n^2 e^{-\frac{4\pi n}{3} y} + M n^2 e^{-2\pi n y}. \tag{6.29}$$

For the moment, we concentrate on the case $n \geq 2$. Since $y \geq \sqrt{3}/2$ for any point on C, we may evaluate this right hand side of (6.29) to get

$$|T_n J e^{-2\pi n y} - 2\cos(2\pi n x)| \leq 1.12. \tag{6.30}$$

This formula is valid for any point on the arc C and $n \geq 2$.

Equation (6.30) is an important result. It places strong constraints on the location of the zeros of $T_n J$. To see this, note that on each of the n intervals I_m defined by $\frac{m}{n} < x < \frac{m+1}{n}$, $m = 0, \ldots, n-1$, the function $2\cos(2\pi n x)$ varies monotonically from -2 to 2. The right hand side of (6.30), however, is less than 2. This implies that $T_n J$ must have n distinct zeroes on C, with one in each interval I_m. As $n \to \infty$, the lengths of these intervals vanish and the zeroes of $T_n J$ become dense on C. Finally, note that, as $T_n J$ is an nth order polynomial in J, it has only n zeroes on D. In particular, it has no additional zeroes beyond those described above.

We have only proven the bound (6.30) for $n \geq 2$. For our application, we also need the $n = 0$ and $n = 1$ cases, which we will consider separately. For $n = 0$, we define $T_0 J = 1$. So the bound (6.30) is trivial. For $n = 1$, we have $T_1 J = J$. In this case we have the slightly weaker bound

$$|J(\tau) e^{-2\pi y} - 2\cos(2\pi x)| \leq 1.22 \tag{6.31}$$

for points on C. This is straightforward to verify numerically, although it may also be understood analytically. The point is that the function $J(\tau)e^{-2\pi y} - 2\cos(2\pi x)$ is monotonic along the arc C. So the value of this function on C is bounded by its values at the endpoints $\tau = i$ and $\tau = 1/2 + \sqrt{3}/2 i$. Using $J(i) = 984$ and $J(1/2 + \sqrt{3}/2 i) = -744$ we may compute the values of the function at these endpoints, giving the bound (6.31).

6.3. Zeroes of Modular Invariant Partition Functions

We are now ready for the general case of a modular invariant partition function at level k:

$$Z_k(\tau) = \sum_{\Delta=-k}^{\infty} F_\Delta q^\Delta = \sum_{\Delta=-k}^{0} F_\Delta T_{|\Delta|} J(\tau) \tag{6.32}$$

where the F_Δ are non-negative integers and $F_{-k} = 1$. As with $J(\tau)$, the fact that $Z_k(\tau)$ has real coefficients and is modular-invariant means that it is real on the imaginary τ axis as well as on the boundary of the fundamental region ∂D.

In the previous section, we proved the estimate (6.30) for $T_n J$ on C with $n \geq 2$. Along with the estimate (6.31) on C, this implies that for all $n \geq 0$

$$T_n J = e^{2\pi n y}(2\cos(2\pi n x) + E_n) \tag{6.33}$$

where E_n is an error term obeying

$$|E_n| < 1.22. \tag{6.34}$$

By simply adding up the inequalities (6.33) for $n = 0, \ldots, k$, with coefficients F_{-n}, and using the fact that $F_{-k} = 1$, we get a a similar estimate for Z_k:

$$Z_k e^{-2\pi k y} - 2\cos(2k\pi x) = E_k + \sum_{\Delta=-k+1}^{0} F_\Delta e^{-2\pi(k+\Delta)y}\left(2\cos(2\pi\Delta x) + E_{|\Delta|}\right). \tag{6.35}$$

To bound the location of the zeroes of Z_k, we must show that the right hand side is less than 2. Then, as in the previous section, the zeroes of Z_k will lie on C and become dense in the large k limit.

For this to be the case, F_Δ must not grow too quickly with Δ. For example, assume that

$$F_\Delta < A e^{2\pi\alpha(\Delta+k)} \qquad \text{for } -k \leq \Delta \leq 0 \tag{6.36}$$

where α and A are positive constants. In this case

$$\begin{aligned} \sum_{\Delta=-k+1}^{0} F_\Delta e^{-2\pi(k+\Delta)y} &< A \sum_{\Delta=-k+1}^{0} e^{2\pi(k+\Delta)(\alpha-y)} \\ &< A\sum_{n=1}^{k} e^{-2\pi n(\alpha-\sqrt{3}/2)} < \frac{A}{e^{2\pi(\alpha-\sqrt{3}/2)}-1}. \end{aligned} \tag{6.37}$$

In the second line, we have used the fact that $y > \sqrt{3}/2$ on C. Since

$$|2\cos(2\pi\Delta x) + E_{|\Delta|}| \leq 2 + |E_{|\Delta|}| < 3.22, \tag{6.38}$$

it follows from (6.34) that

$$|Z_k e^{-2\pi k y} - 2\cos(2k\pi x)| \leq |E_k| + 3.22\frac{A}{e^{2\pi(\sqrt{3}/2-\alpha)}-1}, \tag{6.39}$$

which is less than 2 for certain values of A and α. In particular, using $|E_k| < 1.22$ and setting $A = 1$, we find that the right hand side is less than 2 provided that

$$\alpha \leq 0.61 \tag{6.40}$$

We have used several rather conservative estimates, so it is quite possible that the zeroes of Z_k condense on C even when the coefficients F_Δ do not satisfy the precise bound given above.

It is easy to check that this bound is indeed satisfied for the extremal partition functions of [13]. In fact, the coefficients $Z_k(\Delta)$ grow much more slowly than the required behavior, indicating that many more primary fields could be added without spoiling the existence and nature of the phase transition.

7. Supergravity Partition Functions

Our goal here will be to repeat the analysis of Secs. 2 and 3 for supergravity. We will consider only the basic case of $\mathcal{N} = 1$ supergravity, in which the symmetry group $SL(2,\mathbb{R}) \times SL(2,\mathbb{R})$ of AdS_3 is replaced by $OSp(1|2) \times OSp(1|2)$, where $OSp(1|2)$ is a supergroup whose bosonic part is $Sp(2,\mathbb{R}) = SL(2,\mathbb{R})$. The boundary CFT has $(1,1)$ supersymmetry, that is, $\mathcal{N} = 1$ supersymmetry for both left- and right-movers.

7.1. Formalism

In supergravity, there are a few closely related choices of possible partition function. We can compute either $\mathrm{Tr}\,\exp(-\beta H - i\theta J)$ or $\mathrm{Tr}\,(-1)^F \exp(-\beta H - i\theta J)$. And we can compute this trace in either the Neveu-Schwarz (NS) or the Ramond (R) sector. As usual, we attempt to compute these partition functions by summing over three-manifolds M that are locally AdS_3 and whose conformal boundary Σ is a Riemann surface of genus 1. The four possible partition functions (NS or R, with or without an insertion of $(-1)^F$) correspond to the four spin structures[8] on Σ. The three that correspond to odd spin structures are permuted by the action of $SL(2,\mathbb{Z})$, as we further discuss below.

[8] In fact, using a discrete R-symmetry that appears to be present in the boundary CFT, we could choose different spin structures for left- and right-movers on Σ, but we will not consider this generalization.

Once we pick the spin structure on Σ, we then sum over choices of M such that the given spin structure on Σ does extend over M. For example, what spin structure on Σ is compatible with taking $M = M_{0,1}$ to be the three-manifold related to perturbative excitations of AdS_3?

In $M_{0,1}$, the "spatial" circle on Σ is contractible. This means that the NS spin structure on the spatial circle extends over $M_{0,1}$ and the R spin structure does not. Hence $M_{0,1}$ contributes to traces in the NS sector, not the R sector.

What is the spectrum of thermal excitations in the NS sector? In ordinary gravity, the thermal excitations of left-movers are obtained by acting on the ground state $|\Omega\rangle$ with Virasoro generators L_{-n}, $n \geq 2$. When the boundary theory has $\mathcal{N} = 1$ supersymmetry, we can also act on $|\Omega\rangle$ with superconformal generators $G_{-n+1/2}$, $n \geq 2$. (We recall that $G_{-1/2}|\Omega\rangle = 0$.) Writing $-k^*/2$ for the ground state energy, the partition function of left-moving excitations is therefore

$$q^{-k^*/2} \prod_{n=2}^{\infty} \frac{1 + q^{n-1/2}}{1 - q^n}. \tag{7.1}$$

Including both left- and right-moving excitations, the contribution of $M_{0,1}$ to $F(q, \bar{q}) = \mathrm{Tr}_{\mathrm{NS}} \exp(-\beta H - i\theta J)$ is

$$F_{0,1} = \left| q^{-k^*/2} \prod_{n=2}^{\infty} \frac{1 + q^{n-1/2}}{1 - q^n} \right|^2. \tag{7.2}$$

This formula can be justified exactly as in Sec. 2.2 for the bosonic case.

The complete function F, or more precisely the sum of known contributions to it, is evaluated by summing over all those modular images of $M_{0,1}$ over which the relevant spin structure extends. We can represent the four spin structures on the two-torus Σ by a column vector

$$\begin{pmatrix} \mu \\ \nu \end{pmatrix}, \tag{7.3}$$

where μ and ν represent respectively the fermion boundary conditions in the "time" and "space" directions on the two-torus Σ. We consider μ and ν to be valued in $\frac{1}{2}\mathbb{Z}/\mathbb{Z}$, taking the values $1/2$ for antiperiodic (NS) boundary conditions and 0 for periodic (R) ones. An element of $SL(2, \mathbb{Z})$ acts on the spin structures by

$$\begin{pmatrix} \mu \\ \nu \end{pmatrix} \to \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} \mu \\ \nu \end{pmatrix}. \tag{7.4}$$

The NS partition function $F(\tau) = \mathrm{Tr}_{\mathrm{NS}}\, \exp(-\beta H - i\theta J)$ corresponds to $\mu = \nu = 1/2$. This condition is invariant under the subgroup of $SL(2,\mathbb{Z})$ characterized by saying that $c+d$ and $a+b$ are both odd. If $c+d$ is odd, we can make $a+b$ odd by adding to (a,b) a multiple of (c,d). F, or at least the sum of known contributions to it, can therefore be computed by summing $F_{0,1}$ over modular images with $c+d$ odd:

$$F(\tau) = \sum_{c,d|c+d\,\mathrm{odd}} F_{c,d}(\tau) \tag{7.5}$$

or equivalently

$$F(\tau) = \sum_{c,d|c+d\,\mathrm{odd}} F_{0,1}((a\tau+b)/(c\tau+d)). \tag{7.6}$$

For any given c, d such that $c+d$ is odd, we pick a and b such that $ad-bc=1$ and $a+b$ is odd. Because $F_{0,1}(\tau)$ is invariant under $\tau \to \tau+2$, the summand in (7.6) does not depend on this choice.

It is also of interest to compute partition functions with other spin structures. However, the three even spin structures on Σ (the ones for which μ and ν are not both zero) are permuted by $SL(2,\mathbb{Z})$ and so the associated partition functions are not really independent functions. If we set $\mu = 0$, $\nu = 1/2$, we get $G(\tau) = \mathrm{Tr}_{\mathrm{NS}}\,(-1)^F \exp(-\beta H - i\theta J)$. The contribution of $M_{0,1}$ to this partition function is obtained by reversing the sign of all fermionic contributions in (7.2):

$$G_{0,1}(\tau) = \left| q^{-k^*/2} \prod_{n=2}^{\infty} \frac{1-q^{n-1/2}}{1-q^n} \right|^2 . \tag{7.7}$$

The subgroup of $SL(2,\mathbb{Z})$ that preserves the conditions $\mu = 0$, $\nu = 1/2$ is characterized by requiring that b should be even, which implies that a and d are odd. Hence

$$G(\tau) = \sum_{c,d|d\,\mathrm{odd}} G_{c,d} = \sum_{c,d|d\,\mathrm{odd}} G_{0,1}((a\tau+b)/(c\tau+d)), \tag{7.8}$$

where for given c, d, we pick a, b so that $ad-bc=1$ and b is even. A modular transformation $T : \tau \to \tau+1$ exchanges $(\mu,\nu) = (0,1/2)$ with $(\mu,\nu) = (1/2,1/2)$, so in particular

$$F(\tau) = G(\tau+1) = F(\tau+2), \tag{7.9}$$

and the summand in (7.8) does not depend on the choice of a, b.

The Ramond partition function $K = \mathrm{Tr}_{\mathrm{R}} \exp(-\beta H - i\theta J)$ is computed from $\mu = 1/2$, $\nu = 0$, so

$$K(\tau) = G(-1/\tau). \tag{7.10}$$

This completes our characterization of three of the four partition functions. Finally, in any supersymmetric theory with discrete spectrum, the fourth partition function $I = \mathrm{Tr}_{\mathrm{R}}(-1)^F \exp(-\beta H - i\theta J)$ is an integer, independent of β and θ, since it can be interpreted as the index of a supersymmetry generator. It must be computed using the odd spin structure, the one with $\mu = \nu = 0$. In three-dimensional gravity, assuming that the partition function can be computed by summing over smooth three-geometries, I vanishes, since the odd spin structure does not extend over any three-manifold with boundary Σ.

General Structure

Some simple but important facts follow from (7.9). Since $F(\tau) = \mathrm{Tr}_{\mathrm{NS}} \exp(-\beta H - i\theta J)$ is invariant under $\tau \to \tau + 2$, it follows that all eigenvalues of $J = L_0 - \bar{L}_0$ take values in $\mathbb{Z}/2$. The transformation $\tau \to \tau + 1$ acts as 1 or -1 on states with integer or half-integer J; thus it acts as $(-1)^{2J}$. On the other hand, if we insert a factor of $(-1)^F$ in the trace, we get $G(\tau) = \mathrm{Tr}_{\mathrm{NS}}\,(-1)^F \exp(-\beta H - i\theta J)$. Since $G(\tau) = F(\tau + 1)$, the operator $(-1)^F$ is equivalent to $(-1)^{2J}$. In other words, states of integer or half-integer J are bosonic or fermionic, respectively. The exact theory (to the extent that it can be reconstructed from the sum over smooth classical geometries) inherits this property from the perturbative spectrum of Brown-Henneaux excitations.

The general form of $G(\tau)$, assuming that it has a Hilbert space interpretation, should therefore be

$$G(\tau) = \sum_{j \in \mathbb{Z}} \sum_n a_{n,j} \exp(-\beta E_{n,j} + i\theta j) - \sum_{j \in \mathbb{Z}+1/2} \sum_n a_{n,j} \exp(-\beta E_{n,j} + i\theta j). \tag{7.11}$$

Here $E_{n,j}$, $n = 1, 2, 3, \ldots$, are the energy eigenvalues for states of angular momentum j, and $a_{n,j}$ is the number of states of energy $E_{n,j}$. The $a_{n,j}$ are positive integers; the minus sign preceding the second term in (7.11) reflects the relation $(-1)^{2J} = (-1)^F$.

7.2. The Computation

We we will now compute the partition functions of the previous section by summing over smooth geometries. We will consider the partition function $G(\tau) = \mathrm{Tr}_{\mathrm{NS}}\,(-1)^F \exp(-\beta H - i\theta J)$, as it is technically the simplest; the two other non-zero partition functions are then given by (7.9) and (7.10).

From (7.7) and (7.8) we have

$$G(\tau) = \sum_{c,d|\ d\ \text{odd}} \left(|q|^{-k^*} \prod_{n\geq 2} \left| \frac{1-q^{n-1/2}}{1-q^n} \right|^2 \right) \Bigg|_\gamma . \tag{7.12}$$

To understand the modular transformation properties of this sum, it is useful to rewrite the infinite product in terms of Dedekind eta functions. Using the identities

$$\prod_{n=2}^{\infty} (1-q^n)^{-1} = \frac{q^{1/24}(1-q)}{\eta(\tau)} \tag{7.13}$$

and

$$\prod_{n=2}^{\infty} (1-q^{n-1/2}) = \frac{q^{1/48}}{(1-q^{1/2})} \frac{\eta(\tau/2)}{\eta(\tau)} \tag{7.14}$$

this may be written as

$$G(\tau) = \sum \left(|q|^{-k^*+3/24} |1+q^{1/2}|^2 \frac{|\eta(\tau/2)|^2}{|\eta(\tau)|^4} \right) \Bigg|_\gamma . \tag{7.15}$$

We may now extract these Dedekind eta functions from the sum, using the fact that $\sqrt{\mathrm{Im}\ \tau}\ |\eta(\tau)|^2$ is modular invariant:

$$\begin{aligned} G(\tau) &= \frac{|\eta(\tau/2)|^2}{y^{1/2}|\eta(\tau)|^4} \sum \left(y^{1/2} |q|^{-k^*+3/24} |1-q^{1/2}|^2 \right)_\gamma \\ &= \frac{|\eta(\tau/2)|^2}{y^{1/2}|\eta(\tau)|^4} \big(\widehat{E}(k^*-3/24,0) + \widehat{E}(k^*+1-3/24,0) + \\ &\qquad \widehat{E}(k^*+1/2-3/24,1/2) + \widehat{E}(k^*+1/2-3/24,-1/2) \big) \end{aligned} . \tag{7.16}$$

In the second line we have defined

$$\widehat{E}(\kappa,\mu) = \sum_{c,d|\ d\ \text{odd}} \frac{y^{1/2}}{|c\tau+d|} \exp\left\{ 2\pi\kappa\ \mathrm{Im}\ \gamma\tau + 2\pi i\mu\ \mathrm{Re}\ \gamma\tau \right\} . \tag{7.17}$$

This is the supersymmetric version of the Poincaré series studied in Sec. 3.

This sum is divergent, for the same reasons described in Sec. 3.1. In particular, at large c and d the exponential approaches one and we are left with the linearly divergent sum $\sum_{c,d} |c\tau+d|^{-1}$. The sum may be regularized by considering the more general Poincaré series

$$\widehat{E}(s,\kappa,\mu) = \sum_{c,d|\ d\ \text{odd}} \frac{y^s}{|c\tau+d|^{2s}} \exp\left\{2\pi\kappa\ \text{Im}\ \gamma\tau + 2\pi i\mu\ \text{Re}\ \gamma\tau\right\} \tag{7.18}$$

as an analytic function of s and taking $s \to 1/2$. This regularization scheme can be justified on physical grounds, following the same line of argument presented in Sec. 3.1.

We will now proceed to compute the Fourier coefficients of the sum (7.18). Our calculation is quite similar to that done in Sec. 3. The only differences are that in (7.18), κ and μ are allowed to be half-integer, and that we are summing over c and d such that $(c,d) = 1$ and d is odd. These two conditions on c and d can be combined in to the single condition $(2c,d) = 1$.

We start by letting $d = d' + 2nc$, where $d' = d \mod 2c$. The sum in (7.18) can be written as a sum over c, d', and n:

$$E(s,\kappa,\mu) = y^s e^{2\pi(\kappa y + i\mu x)} + \sum_{c>0} \sum_{d'\in\mathbb{Z}/2c\mathbb{Z}} \sum_{n\in\mathbb{Z}} f(c,d',n) \tag{7.19}$$

where

$$f(c,d',n) = \frac{y^s}{|c(\tau+2n)+d'|^{2s}} \exp\left\{\frac{2\pi\kappa y}{|c(\tau+2n)+d'|^2} + 2\pi i\mu\ \left(\frac{a}{c} - \frac{cx+d}{c|c(\tau+2n)+d'|^2}\right)\right\}. \tag{7.20}$$

We will now apply the Poisson summation formula to the sum over n, as we did for the bosonic partition function in Sec. 3.2. First, we must compute the Fourier transform

$$\begin{aligned}\widehat{f}(c,d',\widehat{n}) &= \int_{-\infty}^{\infty} dn\ e^{2\pi i n\widehat{n}} f(c,d',n) \\ &= \frac{1}{2}\exp\left(2\pi i\frac{(2\mu)a-\widehat{n}d'}{2c} - \pi i\widehat{n}x\right)\int_{-\infty}^{\infty} dt\ e^{\pi i\widehat{n}t}\left(\frac{y}{c^2(t^2+y^2)}\right)^s \exp\left\{\frac{2\pi(\kappa\, y - i\mu t)}{c^2(t^2+y^2)}\right\}.\end{aligned} \tag{7.21}$$

We have written the integral in terms of a shifted integration variable $t = 2n + x + \frac{d'}{c}$.

Recall that, as described after equation (7.7), we must chose our integers a and b such that b is even. The determinant condition $ad-bc = 1$ therefore implies that $ad' = 1 \mod 2c$. Since $2\mu \in \mathbb{Z}$, this implies that the exponential prefactor in (7.21) is a function only of $d' = d \mod 2c$. We may therefore extract the d' dependence of the sum (7.18) into the Kloosterman sum $S(-\widehat{n}, 2\mu; 2c)$, as defined in (3.21). The integral appearing in (7.21) is

precisely that defined in (3.24). The Fourier cofficients of the Poincaré series (7.18) are therefore given by

$$\widehat{E}(s,\kappa,\mu) = y^s e^{2\pi(\kappa y + i\mu x)} + \sum_{\widehat{n}} e^{\pi i \widehat{n} x} \widehat{E}_{\widehat{n}}(s,\kappa,\mu) \tag{7.22}$$

where

$$\widehat{E}_{\widehat{n}}(s,\kappa,\mu) = \frac{1}{2}\sum_{m=0}^{\infty} I_{m,\widehat{n}/2}(s,\kappa,\mu)\; y^{1-m-s} \left(\sum_{c=1}^{\infty} c^{-2(m+s)} S(-\widehat{n}, 2\mu; 2c)\right) \tag{7.23}$$

is defined in terms of the integrals (3.24).

We will now restrict our attention to the $\widehat{n}=0$ case. First consider $\mu = 0$. In this case, the integrals were given in (3.26). To do the sum over c, note first that the Kloosterman sum $S(0,0,2c)$ is equal to Euler's totient function $\phi(2c)$. The sum over c is

$$\sum_{c>0} c^{-2(m+s)} S(0,0,2c) = \sum_{c>0} c^{-2(m+s)}\phi(2c) = \frac{2^{2(m+s)}}{2^{2(m+s)}-1}\frac{\zeta(2(m+s)-1)}{\zeta(2(m+s))}. \tag{7.24}$$

To prove this formula, we first evaluate the sum $\sum_{c\ \mathrm{odd}} c^{-\sigma}\phi(c)$, where we have defined $\sigma = 2(m+s)$. Multiplying this sum by

$$\sum_{n\ \mathrm{odd}} n^{-\sigma} = (1-2^{-\sigma})\zeta(\sigma) \tag{7.25}$$

gives

$$\begin{aligned}(1-2^{-\sigma})\zeta(\sigma)\sum_{c\ \mathrm{odd}} c^{-\sigma}\phi(c) &= \sum_{c,n\ \mathrm{odd}} (cn)^{-\sigma}\phi(c) \\ &= \sum_{m\ \mathrm{odd}} m^{-\sigma}\left(\sum_{c|m}\phi(c)\right) \\ &= \sum_{m\ \mathrm{odd}} m^{-\sigma+1} = (1-2^{1-\sigma})\zeta(\sigma-1)\end{aligned} \tag{7.26}$$

In the second line we defined the new variable $m = cn$ to be summed over, and in the third line we have used the basic identity for the totient function $\sum_{c|m}\phi(c) = m$. We thus have

$$\sum_{c\ \mathrm{odd}} c^{-\sigma}\phi(c) = \frac{2^\sigma - 2}{2^\sigma - 1}\frac{\zeta(\sigma-1)}{\zeta(\sigma)}. \tag{7.27}$$

The sum over even c may be done by recalling the identity that was proved in Sec. 3 by similar methods:

$$\sum_{c>1} c^{-\sigma}\phi(\sigma) = \frac{\zeta(\sigma-1)}{\zeta(\sigma)} \tag{7.28}$$

to get

$$\sum_{c \text{ even}} c^{-\sigma}\phi(c) = \frac{1}{2^\sigma - 1}\frac{\zeta(\sigma-1)}{\zeta(\sigma)}. \tag{7.29}$$

But

$$\sum_{c \text{ even}} c^{-\sigma}\phi(c) = 2^{-\sigma}\sum_{c>0} c^{-\sigma}\phi(2c). \tag{7.30}$$

Combining the last two formulas, we arrive at (7.24).

As in the bosonic case, we must be careful when taking $s \to 1/2$. For $m = 0$ (and $\widehat{n} = 0$), the sum (7.24) vanishes at $s = 1/2$, whereas the integral $I_{0,0}$ has a pole at $m = 0$, $s = 1/2$, as we see in (3.26). The product of the two factors is finite; in fact, $\Gamma(s-1/2)/\zeta(2s) \to 2$ at $s \to 1/2$. The $m > 0$ terms are finite without any such subtleties. Evaluating the first three terms in the expansion of (7.18) gives

$$\widehat{E}_0(1/2,\kappa,0) = -y^{1/2} + \left(\frac{8\pi^3}{21\zeta(3)}\kappa\right) y^{-1/2} + \left(\frac{64\pi^6}{4185\zeta(5)}\kappa^2\right) y^{-3/2} + \mathcal{O}(y^{-5/2}). \tag{7.31}$$

Let us now consider the $\mu = \pm 1/2$ terms. In this case the integrals are only slightly different from those done in Sec. 3, which were evaluated at $\mu = \pm 1$. For $m = 0$, we find that

$$I_{0,0}(s,\kappa,\pm 1/2) = \sqrt{\pi}\frac{\Gamma(s-1/2)}{\Gamma(s)}. \tag{7.32}$$

This is the analog of equation (3.28) used in the computation of the bosonic partition function. For $m > 0$ the integrals are complicated hypergeometric functions of the sort written down in (3.27). At $s = 1/2$ these hypergeometric functions simply to give

$$I_{m,0}(1/2,\kappa,\pm 1/2) = \frac{2^{1-m}\pi^{m+1/2}}{m\Gamma(m+1/2)}T_m(2\kappa). \tag{7.33}$$

This is the analog of equation (3.42) used in the bosonic case.

To do the sum over c, we use the fact that the Kloosterman sum $S(0,\pm 1,2c)$ is equal to the Möbius function $\mu(2c)$. The sum over c is given by

$$\sum_{c>0} c^{-2(m+s)} S(0,\pm 1,2c) = \sum_{c} c^{-2(m+s)}\mu(2c) = -\frac{2^{2(m+s)}}{(2^{2(m+s)}-1)\zeta(2(m+s))}. \tag{7.34}$$

We can obtain this formula as follows. We first compute the sum $\sum_{c \text{ odd}} c^{-\sigma}\mu(c)$. Multiplying this sum by (7.25) we get

$$\begin{aligned}(1-2^{-\sigma})\zeta(\sigma)\sum_{c \text{ odd}} c^{-\sigma}\mu(c) &= \sum_{c,n \text{ odd}} (cn)^{-\sigma}\mu(c) \\ &= \sum_{m \text{ odd}} m^{-\sigma}\left(\sum_{c|m}\mu(c)\right) \\ &= 1\end{aligned} \tag{7.35}$$

In the second line we have let $m = cn$ and in the third line we have used the fact that for any m, $\sum_{c|m} \mu(c) = \delta_{m,1}$. The sum over even c may be done by recalling the similar identity from Sec. 3

$$\sum_{c>1} c^{-\sigma} \mu(\sigma) = \frac{1}{\zeta(\sigma)} \tag{7.36}$$

to get

$$\sum_{c \text{ even}} c^{-\sigma} \phi(c) = -\frac{1}{2^\sigma - 1} \frac{1}{\zeta(\sigma)}, \tag{7.37}$$

which is equivalent to (7.34).

At $s = 1/2$, we again must be careful to cancel the zero in (7.34) at $m = 0$ against a pole in (7.32), using the fact that $\Gamma(s - 1/2)/\zeta(2s) \to 2$ as $s \to 1/2$. Including the next two terms in the series, we find

$$E_0(1/2, \kappa, \pm 1/2) = -2y^{1/2} - \left(\frac{16\pi}{7\zeta(3)}\kappa\right) y^{-1/2} - \left(\frac{16\pi^2}{93\zeta(5)}(8\kappa^2 - 1)\right) y^{-3/2} + \mathcal{O}(y^{-5/2}) \tag{7.38}$$

Putting this all together gives the following expansion for the partition function

$$\begin{aligned} G(\tau) =& G_{0,1}(\tau) + \frac{|\eta(\tau/2)|^2}{|\eta(\tau)|^4}\Big(-6 + \frac{(6 + 16k^*)(\pi^3 - 6\pi)}{21\zeta(3)} y^{-1} \\ & - \frac{4\pi^2(2880k^{*2} - 16k^*(2\pi^4 - 135) + 45 - 12\pi^4)}{4185\zeta(5)} y^{-2} + \mathcal{O}(y^{-3})\Big). \end{aligned} \tag{7.39}$$

As in the bosonic case, a Hilbert space interpretation is precluded both by the fact that the coefficient of the leading correction to $G_{0,1}(\tau)$ is negative and by the fact that there are additional corrections involving powers of $1/y$. A Hilbert space interpretation would require the leading coefficient to be positive in view of (7.11); the leading correction to the Brown-Henneaux spectrum arises for an integer value of the angular momentum, namely $j = \widehat{n} = 0$.

We have so far considered just the $\widehat{n} = 0$ case. The discussion for the $\widehat{n} \neq 0$ case will be qualitatively similar to the bosonic case discussed in Sec. 3.4. The integrals can be evaluated explicitly – they are the same ones we considered in Sec. 3.4 – and contribute terms to the NS sector partition function which fall off exponentially at large y. As in the bosonic case, the sums over c cannot be expressed in terms of elementary number theoretic functions but are related to an appropriately defined Selberg zeta function. In particular, they should be related to the Selberg zeta function associated to the congruence subgroup described in Sec. 7.1, where b is constrained to be even. As in Sec. 3.4, this should provide

a finite regularization of the sum because the associated Laplacian does not have a discrete eigenfunction with eigenvalue 1/4.

We conclude this section by emphasizing a few possible interpretations of the result (7.39) for the supergravity partition function $G(\tau)$. As in the bosonic case, this partition function does not have the structure of a Hilbert space trace $\mathrm{Tr}_{\mathrm{NS}}\ (-1)^F e^{-\beta H - i\theta J}$. One possible interpretation of this result is that three-dimensional pure supergravity does not exist as a quantum theory, and that additional degrees of freedom must be included in order to render the theory sensible. A second possible implication is that the pure supergravity theory exists, but contains long strings, similar to those described in Sec. 4.1. A third possibility is that additional complex saddle points must be included in the sum over geometries, as described for the bosonic theory in Sec. 4.2. This might lead to a sum over two copies of the modular group, resulting in a holomorphic partition function which might coincide with the extremal supergravity partition functions described in [13].

7.3. Phase Transitions

In this section we will comment briefly on the supergravity generalization of the results of Sec. 6. In particular, we ask whether, for holomorphically factorized supergravity partition functions, Hawking-Page phase transitions occur via a condensation of zeroes along curves in the complex temperature plane. Numerical evidence described to us by M. Kaneko suggests that this may be the case.

Let us first ask what the phase diagram of three-dimensional supergravity should be. We will consider the NS partition function $F(\tau) = \mathrm{Tr}_{\mathrm{NS}}\ e^{-\beta H - i\theta J}$. As described in Sec. 7.1, $F(\tau)$ is invariant under the subgroup of $SL(2,\mathbb{Z})$ defined by the condition that $a+b$ and $c+d$ are both odd; this group is sometimes called Γ_θ. The tessellation of the upper half plane $\mathcal{H}$ by fundamental domains of Γ_θ is shown in Fig. 4a).

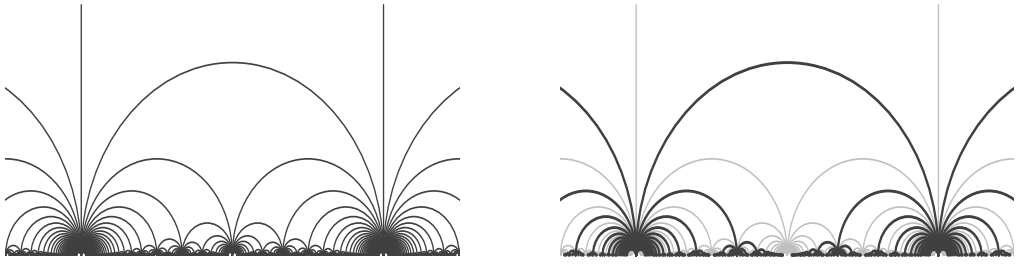


Figure 4: a) The tessellation of the upper half τ-plane by fundamental domains of Γ_θ. b) The subtessellation of the upper half plane corresponding to the coset $\Gamma_\theta/\mathbb{Z}$ (dark lines). This is the phase diagram of three-dimensional supergravity. The tessellation by Γ_θ (light lines) is shown to guide the eye.

Within this modular group Γ_θ, there is a subgroup isomorphic to $\mathbb{Z}$ generated by the translation $\tau \to \tau + 2$. This $\mathbb{Z}$ is the intersection of Γ_θ with the mapping class group of the handlebody $M_{0,1}$, so modular transformations in $\mathbb{Z}$ do not generate new handlebody contributions to the supergravity partition function. There is a unique saddle point contribution to $F(\tau)$ for each element of the coset $\Gamma_\theta/\mathbb{Z}$. The subtessellation of the upper half plane $\mathcal{H}$ associated to this coset $\Gamma_\theta/\mathbb{Z}$ is shown in Fig. 4b). As one moves between tiles in this diagram, we expect a first order Hawking-Page phase transition as different saddles become dominant. So the subtessellation depicted in Fig. 4b) should be interpreted as the phase diagram of three dimensional supergravity.

In [13], holomorphically factorized partition functions for three-dimensional gravity were considered. Numerically, it appears that[9] the zeroes of these partition functions lie on the dark curves in Figure 4b), and become dense on these curves in the large k limit [40]. This indicates that the Hawking-Page transition in supergravity occurs by the mechanism of Lee-Yang and Fischer, as in the bosonic case described in Sec. 6. It may be possible to prove this by extending the analytic arguments of [37] and Sec. 6.

Acknowledgments

We are very grateful to R. Canary, D. Gaiotto, M. Kaneko, J. Maldacena, M. Mirzakhani, R. Rabadan, N. Seiberg, X. Yin, and especially P. Sarnak for useful conversations.

[9] Here we assume the numbers of Ramond and Neveu-Schwarz primary states at the black hole threshold to be small. In [13], there was no natural way to determine these numbers.

References

[1] S. Deser and R. Jackiw, "Three-Dimensional Cosmological Gravity: Dynamics Of Constant Curvature," Annals Phys. **153**, 405 (1984).

[2] J. D. Brown and M. Henneaux, "Central Charges In The Canonical Realization Of Asymptotical Symmetries: An Example From Three-Dimensional Gravity," Commun. Math. Phys. **104** (1986) 207-226.

[3] M. Bañados, C. Teitelboim, and J. Zanelli, "The Black Hole In Three-Dimensional Spacetime," Phys. Rev. Lett. **69** (1992) 1849-1851, hep-th/9204099.

[4] M. Bañados, M. Henneaux, C. Teitelboim and J. Zanelli, "Geometry of the (2+1) Black Hole," Phys. Rev. D **48**, 1506 (1993) [arXiv:gr-qc/9302012].

[5] J. Maldacena, "The Large N Limit Of Superconformal Field Theories And Supergravity," Adv. Theor. Math. Phys. **2** (1998) 231-252, hep-th/9711200.

[6] A. Strominger, "Black Hole Entropy From Near Horizon Microstates," JHEP 9802:009 (1998), hep-th/9712251.

[7] S. Carlip, "Conformal Field Theory, $(2+1)$-Dimensional Gravity, and the BTZ Black Hole," gr-qc/0503022.

[8] G. W. Gibbons, S. W. Hawking and M. J. Perry, "Path Integrals And The Indefiniteness Of The Gravitational Action," Nucl. Phys. B **138**, 141 (1978).

[9] J. M. Maldacena and A. Strominger, "AdS(3) Black Holes and a Stringy Exclusion Principle," JHEP **9812**, 005 (1998) [arXiv:hep-th/9804085].

[10] R. Dijkgraaf, J. Maldacena, G. Moore, and E. Verlinde, "A Black Hole Farey Tail," hep-th/0005003.

[11] D. B. Ray and I. M. Singer, "Analytic Torsion For Complex Manifolds," Annals Math. **98**, 154 (1973).

[12] X. Yin, "Partition Functions of Three-Dimensional Pure Gravity," arXiv:0710.2129 [hep-th].

[13] E. Witten, "Three-Dimensional Gravity Revisited," arXiv:0706.3359 [hep-th].

[14] J. Manschot, "AdS_3 Partition Functions Reconstructed," arXiv:0707.1159 [hep-th].

[15] S. W. Hawking and D. N. Page, "Thermodynamics Of Black Holes In Anti-De Sitter Space," Commun. Math. Phys. **87**, 577 (1983).

[16] S. Carlip and C. Teitelboim, "Aspects Of Black Hole Quantum Mechanics And Thermodynamics In (2+1)-Dimensions," Phys. Rev. D **51**, 622 (1995) [arXiv:gr-qc/9405070].

[17] S. Deser, R. Jackiw and G. 't Hooft, "Three-Dimensional Einstein Gravity: Dynamics Of Flat Space," Annals Phys. **152**, 220 (1984).

[18] A. Achúcarro and P. Townsend, "A Chern-Simons Action for Three-Dimensional anti-De Sitter Supergravity Theories," Phys. Lett. **B180** (1986) 89.

[19] E. Witten, "(2+1)-Dimensional Gravity as an Exactly Soluble System," Nucl. Phys. **B311** (1988) 46.

[20] E. Witten, "Coadjoint Orbits Of The Virasoro Group," Commun. Math. Phys. **114** (1988) 1-53.

[21] A. A. Bytsenko, L. Vanzo and S. Zerbini, "Quantum Correction to the Entropy of the (2+1)-Dimensional Black Hole," Phys. Rev. D **57**, 4917 (1998) [arXiv:gr-qc/9710106].

[22] H. Iwaniec, *Spectral Methods of Automorphic Forms*, American Mathematical Society, 2002.

[23] M. Kleban, M. Porrati and R. Rabadan, "Poincare Recurrences and Topological Diversity," JHEP **0410**, 030 (2004) [arXiv:hep-th/0407192].

[24] J. Maldacena, J. Michelson, and A. Strominger, "Anti de Sitter Fragmentation," JHEP 9902:011 (1999), hep-th/9812073.

[25] N. Seiberg and E. Witten, "The D1/D5 System and Singular CFT," JHEP 9904:017 (1999), hep-th/9903224.

[26] K. Krasnov, "On Holomorphic Factorization In Asymptotically AdS 3D Gravity," Class. Quant.Grav. **20** (2003(4015-4042, hep-th/0109198.

[27] G. Höhn, "Selbstduale Vertexoperatorsuperalgebren und das Babymonster," Ph.D. thesis (Bonn 1995), Bonner Mathematische Schriften 286 (1996), 1-85, English translation, arXiv:0706.0236.

[28] D. Gaiotto and X. Yin, "Genus Two Partition Functions of Extremal Conformal Field Theories," JHEP 0708:029 (2007), arXiv:0707.3437.

[29] X. Yin, "Partition Functions of Three-Dimensional Pure Gravity," arXiv:0710.2129.

[30] M. Gaberdiel, " Constraints on Extremal Self-Dual CFTs," arXiv:0707.4073.

[31] D. Gaiotto, to appear.

[32] I. B. Frenkel, J. Lepowsky, and A. Meurman, *Vertex Operator Algebras And The Monster* (Academic Press, Boston, 1988).

[33] S. Carlip, Class. Quant. Grav. **17**, 4175 (2000) [arXiv:gr-qc/0005017].

[34] C. N. Yang and T.-D. Lee, "Statistical Theory Of Equations Of State And Phase Transitions. I. Theory Of Condensation," Phys. Rev. (1952) **87** 404-9.

[35] M. Fisher, in *Lectures In Theoretical Physics* VIIC (University of Colorado Press, Boulder, 1965).

[36] R. Rankin, "The Zeros of Certain Poincaré Series," Compositio Mathmatica, **46** (1982) 255-272.

[37] T. Asai, M. Kaneko, H. Ninomiya, "Zeros of Certain Modular Functions and an Application," Commentarii Math. Univ. Sancti Pauli, vol. 46-1 (1997) 93-101.

[38] J. D. Bekenstein, "A Universal Upper Bound On The Entropy To Energy Ratio For Bounded Systems," Phys. Rev. D **23**, 287 (1981).

[39] T. Apostol, *Modular Functions and Dirichlet Series in Number Theory*, Springer Verlag, 1990.

[40] M. Kaneko, private communication.

arXiv:0801.4566v2 [hep-th] 21 Jul 2008

CHIRAL GRAVITY IN THREE DIMENSIONS

Wei Li, Wei Song[1] and Andrew Strominger

Center for the Fundamental Laws of Nature
Jefferson Physical Laboratory, Harvard University, Cambridge MA 02138, USA

Abstract

Three dimensional Einstein gravity with negative cosmological constant $-1/\ell^2$ deformed by a gravitational Chern-Simons action with coefficient $1/\mu$ is studied in an asymptotically AdS_3 spacetime. It is argued to violate unitary or positivity for generic μ due to negative-energy massive gravitons. However at the critical value $\mu\ell = 1$, the massive gravitons disappear and BTZ black holes all have mass and angular momentum related by $\ell M = J$. The corresponding chiral quantum theory of gravity is conjectured to exist and be dual to a purely right-moving boundary CFT with central charges $(c_L, c_R) = (0, 3\ell/G)$.

[1]Permanent address: *Institute of Theoretical Physics Academia Sinica, Beijing 100080, China.*

Contents

1 Introduction 1

2 Topologically Massive Gravity 4

2.1 Action 4

2.2 AdS_3 vacuum solution 5

2.3 Chiral gravity 5

3 Gravitons in AdS_3 6

4 Positivity of energy 11

4.1 BTZ black holes 11

4.2 Gravitons 12

5 Conclusion and Discussion 14

1 Introduction

In three dimensions, the Riemann and Ricci tensor both have the same number (six) of independent components. Hence Einstein's equation, with or without a cosmological constant Λ, completely constrains the geometry and there are no local propagating degrees of freedom. At first sight this makes the theory sound too trivial to be interesting. However in the case of a negative cosmological constant, there are asymptotically AdS_3 black hole solutions [1] as well as massless gravitons which can be viewed as propagating

on the boundary. These black holes obey the laws of black hole thermodynamics and have an entropy given by one-quarter the horizon area. This raises the interesting question: what is the microscopic origin of the black hole entropy in these "trivial" theories?

In order to address this question one must quantize the theory. One proposal [2] is to recast it as an $SL(2,\mathbb{R})_L \times SL(2,\mathbb{R})_R$ Chern-Simons gauge theory with $k_L = k_R$. Despite some effort this approach has not given a clear accounting of the black hole entropy (see however [3] for an interesting attempt). So the situation remains unsatisfactory.

More might be learned by deforming the theory with the addition of the gravitational Chern-Simons term with coefficient $\frac{1}{\mu}$ [4] [5]. The resulting theory is known as the topologically massive gravity (TMG) and contains a local, massive propagating degree of freedom, as well as black holes and massless boundary gravitons. The addition of the Chern-Simons term leads to more degrees of freedom because it contains three, rather than just two, derivatives of the metric. It is the purpose of this note to study this theory for negative $\Lambda = -1/\ell^2$. We will argue that the theory is unstable/inconsistent for generic μ: either the massive gravitons or BTZ black holes have negative energy. The exception occurs when the parameters obey $\mu\ell = 1$, at which point several interesting phenomena simultaneously arise:

(i) The central charges of the dual boundary CFT become $c_L = 0$, $c_R = 3\ell/G$.

(ii) The conformal weights as well as the wave function of the massive graviton, generically $\frac{1}{2}(3 + \mu\ell, -1 + \mu\ell)$, degenerate with those of the left-moving weight $(2, 0)$ massless boundary graviton. They are both pure gauge, but the gauge transformation parameter does not vanish at infinity.

(iii) BTZ black holes and all gravitons have non-negative masses. Further the angular momentum is fixed in terms of the mass to be $J = M\ell$.

This suggests the possibility of a stable, consistent theory at $\mu\ell = 1$ which is dual to a holomorphic boundary CFT (i.e. containing only right-moving degrees of freedom) with $c_R = 3\ell/G$. The hope - which remains

to be investigated - is that for a suitable choice of boundary conditions the zero-energy left-moving excitations can be discarded as pure gauge. We will refer to this theory as 3D chiral gravity. As we will review herein, if such a dual CFT exists, and is unitary, an application of the Cardy formula gives a microscopic derivation of the black hole entropy [6] [7] [8].

Related recent work [9]-[16] has considered an alternative deformation of pure 3D gravity, locally described by the $SL(2,\mathbb{R})_L \times SL(2,\mathbb{R})_R$ Chern-Simons gauge theory with $k_L \neq k_R$. This is a purely topological theory with no local degrees of freedom and is not equivalent to TMG. It contains all the subset of solutions of TMG which are Einstien metrics but not the massive gravitons. It is nevertheless possible that the arguments given in [9] (adapted to the case $k_L = 0$ as in [10]) which are quite general apply to the chiral gravity discussed herein. Indeed discrepancies with the semiclassical analysis mentioned in [9]-[16] disappear for the special case $k_L = 0$.[2] Moreover the main assumption of [9] - holomorphic factorization of the partition function - is simply a consistency requirement for chiral gravity because there are only right movers.

The paper is organized as follows. Section 2 gives a brief review of the cosmological TMG and its AdS_3 vacuum solution, and shows that the theory is purely chiral at the special value of $\mu = 1/\ell$. Section 3 describes the linearized gravitational excitations around AdS_3. Section 4 shows how the positivity of energy imposes a stringent constraint on the allowed value of μ. We end with a short summary and discussion of future directions.

[2] Although we do not study the Euclidean theory herein the relation $M\ell = J$ for Lorentzian BTZ black holes in chiral gravity suggests that the saddle point action will be holomorphic.

2 Topologically Massive Gravity

2.1 Action

The action for TMG (topologically massive gravity) with a negative cosmological constant is [17]

$$I = \frac{1}{16\pi G}\int d^3x\sqrt{-g}(R-2\Lambda) + \frac{1}{16\pi G\mu}I_{CS} \tag{1}$$

where I_{cs} is the Chern-Simons term

$$I_{cs} = \frac{1}{2}\int d^3x\sqrt{-g}\epsilon^{\lambda\mu\nu}\Gamma^{\rho}_{\lambda\sigma}[\partial_\mu\Gamma^{\sigma}_{\rho\nu} + \frac{2}{3}\Gamma^{\sigma}_{\mu\tau}\Gamma^{\tau}_{\nu\rho}] \tag{2}$$

and we take Λ negative. We have chosen the sign in front of the Einstein-Hilbert action so that BTZ black holes have positive energy for large μ, while the massive gravitons will turn out to have negative energy for this choice.[3] The equation of motion is

$$\mathcal{G}_{\mu\nu} + \frac{1}{\mu}C_{\mu\nu} = 0\ , \tag{3}$$

where $\mathcal{G}_{\mu\nu}$ is the cosmological-constant-modified Einstein tensor:

$$\mathcal{G}_{\mu\nu} \equiv R_{\mu\nu} - \frac{1}{2}g_{\mu\nu}R + \Lambda g_{\mu\nu} \tag{4}$$

and $C_{\mu\nu}$ is the Cotton tensor

$$C_{\mu\nu} \equiv \epsilon_{\mu}{}^{\alpha\beta}\nabla_\alpha(R_{\beta\nu} - \frac{1}{4}g_{\beta\nu}R). \tag{5}$$

Einstein metrics with $\mathcal{G}_{\mu\nu} = 0$ are a subset of the general solutions of (3).

[3] This contrasts with most of the literature which chooses the opposite sign in order that massive gravitons have positive energy (for large μ).

2.2 AdS_3 vacuum solution

TMG has an AdS_3 solution:

$$ds^2 = \bar{g}_{\mu\nu}dx^\mu dx^\nu = \ell^2(-\cosh^2\rho d\tau^2 + \sinh^2\rho d\phi^2 + d\rho^2) \tag{6}$$

where the radius is related to Λ by

$$\ell^{-2} = -\Lambda \tag{7}$$

The Riemann tensor, Ricci tensor and Ricci scalar of the AdS_3 are:

$$\bar{R}_{\mu\alpha\nu\beta} = \Lambda(\bar{g}_{\mu\nu}\bar{g}_{\alpha\beta} - \bar{g}_{\mu\beta}\bar{g}_{\alpha\nu}), \quad \bar{R}_{\mu\nu} = 2\Lambda\bar{g}_{\mu\nu}, \quad \bar{R} = 6\Lambda, \tag{8}$$

The metric (6) has isometry group $SL(2,\mathbb{R})_L \times SL(2,\mathbb{R})_R$. The $SL(2,\mathbb{R})_L$ generators are

$$L_0 = i\partial_u \ , \tag{9}$$

$$L_{-1} = ie^{-iu}\left[\frac{\cosh 2\rho}{\sinh 2\rho}\partial_u - \frac{1}{\sinh 2\rho}\partial_v + \frac{i}{2}\partial_\rho\right] \ , \tag{10}$$

$$L_1 = ie^{iu}\left[\frac{\cosh 2\rho}{\sinh 2\rho}\partial_u - \frac{1}{\sinh 2\rho}\partial_v - \frac{i}{2}\partial_\rho\right] \tag{11}$$

where $u \equiv \tau + \phi, v \equiv \tau - \phi$. The $SL(2,\mathbb{R})_R$ generators $\{\bar{L}_0, \bar{L}_{\pm 1}\}$ are given by the above expressions with $u \leftrightarrow v$. The normalization of the $SL(2,\mathbb{R})$ algebra is

$$[L_0, L_{\pm 1}] = \mp L_{\pm 1}, \quad [L_1, L_{-1}] = 2L_0. \tag{12}$$

The quadratic Casimir of $SL(2,\mathbb{R})_L$ is $L^2 = \frac{1}{2}(L_1L_{-1} + L_{-1}L_1) - L_0^2$. When acting on scalars, $L^2 + \bar{L}^2 = -\frac{\ell^2}{2}\bar{\nabla}^2$ [18].

2.3 Chiral gravity

As shown by Brown and Henneaux [19], quantum gravity on asymptotically AdS_3 spacetimes with appropriate boundary conditions is described by a 2D CFT which lives on the boundary. They computed the total central charge of the CFT and found $c_L + c_R = 3\ell/G$. Simplified calculations were given

in [20] [21] [22]. The difference $c_L - c_R$ corresponds to the diffeomorphism anomaly. In reference [8] [23] it is shown that $c_L - c_R = -\frac{3}{\mu G}$. In summary, we have

$$c_L = \frac{3\ell}{2G}(1 - \frac{1}{\mu\ell}) \qquad c_R = \frac{3\ell}{2G}(1 + \frac{1}{\mu\ell}) \tag{13}$$

In order that both central charges are non-negative, we must have, as was also noticed in [24],

$$\mu\ell \geqslant 1 \tag{14}$$

We note that had we chosen the opposite sign in front of the Einstein-Hilbert action the central charge would be negative. An interesting special case is

$$\mu\ell = 1 \tag{15}$$

which implies

$$c_L = 0, \qquad c_R = \frac{3\ell}{G}. \tag{16}$$

We will refer to this theory as chiral gravity. If the chiral gravity is unitary it can have only right-moving excitations.

3 Gravitons in AdS_3

In this section we describe the linearized excitations around background AdS_3 metric $\bar{g}_{\mu\nu}$. Expanding

$$g_{\mu\nu} = \bar{g}_{\mu\nu} + h_{\mu\nu} \tag{17}$$

with $h_{\mu\nu}$ small, the linearized Ricci tensor and Ricci scalar are [25] [26],

$$R^{(1)}_{\mu\nu} = \frac{1}{2}(-\bar{\nabla}^2 h_{\mu\nu} - \bar{\nabla}_\mu\bar{\nabla}_\nu h + \bar{\nabla}^\sigma\bar{\nabla}_\nu h_{\sigma\mu} + \bar{\nabla}^\sigma\bar{\nabla}_\mu h_{\sigma\nu}) \tag{18}$$

$$R^{(1)} \equiv (R_{\mu\nu}g^{\mu\nu})^{(1)} = -\bar{\nabla}^2 h + \bar{\nabla}_\mu\bar{\nabla}_\nu h^{\mu\nu} - 2\Lambda h. \tag{19}$$

The leading terms in $\mathcal{G}$ and C are

$$\mathcal{G}^{(1)}_{\mu\nu} = R^{(1)}_{\mu\nu} - \frac{1}{2}\bar{g}_{\mu\nu}R^{(1)} - 2\Lambda h_{\mu\nu}, \tag{20}$$

$$C^{(1)}_{\mu\nu} = \epsilon_\mu{}^{\alpha\beta}\bar{\nabla}_\alpha\left(R^{(1)}_{\beta\nu} - \frac{1}{4}\bar{g}_{\beta\nu}R^{(1)} - 2\Lambda h_{\beta\nu}\right). \tag{21}$$

We note $\overline{\mathrm{Tr}}C^{(1)} = 0$ and the Bianchi identity implies $\bar{\nabla}^\mu \mathcal{G}^{(1)}_{\mu\nu} = \bar{\nabla}^\mu C^{(1)}_{\mu\nu} = 0$. The linearized equations of motion are then

$$\mathcal{G}^{(1)}_{\mu\nu} + \frac{1}{\mu} C^{(1)}_{\mu\nu} = 0 \tag{22}$$

Tracing this equation yields $\overline{\mathrm{Tr}}\mathcal{G}^{(1)} = -\frac{1}{2}R^{(1)} = 0$. So the equation of motion becomes

$$\mathcal{G}^{(1)}_{\mu\nu} + \frac{1}{\mu}\epsilon_\mu{}^{\alpha\beta}\bar{\nabla}_\alpha \mathcal{G}^{(1)}_{\beta\nu} = 0 \tag{23}$$

where in terms of $h_{\mu\nu}$

$$\mathcal{G}^{(1)}_{\mu\nu} = \frac{1}{2}(-\bar{\nabla}^2 h_{\mu\nu} - \bar{\nabla}_\mu\bar{\nabla}_\nu h + \bar{\nabla}^\sigma\bar{\nabla}_\nu h_{\sigma\mu} + \bar{\nabla}^\sigma\bar{\nabla}_\mu h_{\sigma\nu}) - 2\Lambda h_{\mu\nu} \tag{24}$$

Now we fix the gauge. We define $\tilde{h}_{\mu\nu} \equiv h_{\mu\nu} - \bar{g}_{\mu\nu}h$, which gives $\tilde{h} = -2h$. Plugging $h_{\mu\nu} = \tilde{h}_{\mu\nu} - \frac{1}{2}\bar{g}_{\mu\nu}\tilde{h}$ into (19) and setting it to zero gives

$$\bar{\nabla}_\mu\bar{\nabla}_\nu\tilde{h}^{\mu\nu} = -\Lambda\tilde{h} \tag{25}$$

Thus, the gauge

$$\bar{\nabla}_\mu\tilde{h}^{\mu\nu} = 0 \tag{26}$$

together with the linearized equation of motion implies tracelessness of $h_{\mu\nu}$: $\tilde{h} = -2h = 0$. This gauge is equivalent to the harmonic plus traceless gauge $\bar{\nabla}_\mu h^{\mu\nu} = h = 0$. Noting that

$$[\bar{\nabla}_\sigma, \bar{\nabla}_\mu]h^\sigma_\nu = \bar{R}^\sigma{}_{\lambda\sigma\mu}h^\lambda_\nu - \bar{R}^\lambda{}_{\nu\sigma\mu}h^\sigma_\lambda = 3\Lambda h_{\mu\nu} - \Lambda h g_{\mu\nu} \tag{27}$$

and imposing the gauge condition, (24) is just

$$\mathcal{G}^{(1)}_{\mu\nu} = \frac{1}{2}(-\bar{\nabla}^2 h_{\mu\nu} + 2\Lambda h_{\mu\nu}\ ) \ . \tag{28}$$

The equation of motion (23) thus becomes

$$(\bar{\nabla}^2 + \frac{2}{\ell^2})(h_{\mu\nu} + \frac{1}{\mu}\epsilon_\mu{}^{\alpha\beta}\bar{\nabla}_\alpha h_{\beta\nu}) = 0 \tag{29}$$

Define three mutually commuting operators $(\mathcal{D}^L, \mathcal{D}^R, \mathcal{D}^M)$:

$$(\mathcal{D}^{L/R})_\mu{}^\beta \equiv \delta_\mu{}^\beta \pm \ell\epsilon_\mu{}^{\alpha\beta}\bar{\nabla}_\alpha, \qquad \text{and} \qquad (\mathcal{D}^M)_\mu{}^\beta \equiv \delta_\mu{}^\beta + \frac{1}{\mu}\epsilon_\mu{}^{\alpha\beta}\bar{\nabla}_\alpha \tag{30}$$

where the meaning of superscripts will become clear presently. The equations of motion (23) can then be written as

$$(\mathcal{D}^L\mathcal{D}^R\mathcal{D}^M h)_{\mu\nu} = 0 \tag{31}$$

where we have used the linearized Bianchi identity. Since the three operators commute equation (31) has three branches of solutions. First, the massive gravitons $h^M_{\mu\nu}$ given by

$$(\mathcal{D}^M h^M)_{\mu\nu} = h^M_{\mu\nu} + \frac{1}{\mu}\epsilon_\mu{}^{\alpha\beta}\bar{\nabla}_\alpha h^M_{\beta\nu} = 0 \tag{32}$$

are solutions special for TMG. The other two branches are massless gravitons which are also solutions of Einstein gravity: $\mathcal{G}^{(1)}_{\mu\nu} = 0$. The left-mover $h^L_{\mu\nu}$ and right-mover $h^R_{\mu\nu}$ have different first order equations of motion:

$$(\mathcal{D}^L h^L)_{\mu\nu} = h^L_{\mu\nu} + \ell\epsilon_\mu{}^{\alpha\beta}\bar{\nabla}_\alpha h^L_{\beta\nu} = 0 \qquad (\mathcal{D}^R h^R)_{\mu\nu} = h^R_{\mu\nu} - \ell\epsilon_\mu{}^{\alpha\beta}\bar{\nabla}_\alpha h^R_{\beta\nu} = 0 \tag{33}$$

Note that the components of (32) and (33) tangent to the AdS_3 boundary relate those components of $h_{\mu\nu}$ to their falloff at infinity. This could be used to directly infer their conformal weights but we will instead just find the full solutions.

Next we solve for the three branches of solutions. Define linear operator $(\tilde{\mathcal{D}}^M)_\mu{}^\beta \equiv \delta_\mu{}^\beta - \frac{1}{\mu}\epsilon_\mu{}^{\alpha\beta}\bar{\nabla}_\alpha$, which commutes with $\mathcal{D}^M$ defined earlier. Applying $\tilde{\mathcal{D}}^M$ on (23), we get a second order equation,

$$(\tilde{\mathcal{D}}^M\mathcal{D}^M\mathcal{G}^{(1)})_{\mu\nu} = (\mathcal{D}^M\tilde{\mathcal{D}}^M\mathcal{G}^{(1)})_{\mu\nu} = -\frac{1}{\mu^2}[\bar{\nabla}^2 - (\mu^2 + 3\Lambda)]\mathcal{G}^{(1)}_{\mu\nu} = 0 \tag{34}$$

where we have used the linearized Bianchi identity. Note that $\tilde{\mathcal{D}}^M\mathcal{G} = 0$ is just the linearized gravitational wave equation if we exchange μ for $-\mu$. So solutions of TMG with both signs of μ are solutions of (34). It can conversely be shown that all solutions of (34) are solutions of TMG for one sign of μ or the other.

Rewrite the Laplacian acting on rank two tensors in terms of the sum of two $SL(2,\mathbb{R})$ Casimirs:

$$\bar{\nabla}^2 h_{\mu\nu} = -[\frac{2}{\ell^2}(L^2 + \bar{L}^2) + \frac{6}{\ell^2}]h_{\mu\nu} \tag{35}$$

$\mathcal{G}^{(1)}_{\mu\nu}$ can be written as

$$\mathcal{G}^{(1)}_{\mu\nu} = [\frac{1}{\ell^2}(L^2 + \bar{L}^2) + \frac{2}{\ell^2}]h_{\mu\nu} \ . \tag{36}$$

Thus (34) becomes

$$[-\frac{2}{\ell^2}(L^2 + \bar{L}^2) - \frac{3}{\ell^2} - \mu^2][\frac{1}{\ell^2}(L^2 + \bar{L}^2) + \frac{2}{\ell^2}]h_{\mu\nu} = 0 \ . \tag{37}$$

This allows us to use the $SL(2,\mathbb{R})_L \times SL(2,\mathbb{R})_R$ algebra to classify the solutions of (34). Consider states with weight $(h, \bar{h})$:

$$L_0|\psi_{\mu\nu}\rangle = h|\psi_{\mu\nu}\rangle, \qquad \bar{L}_0|\psi_{\mu\nu}\rangle = \bar{h}|\psi_{\mu\nu}\rangle. \tag{38}$$

From the explicit form of the generators, we see

$$\psi_{\mu\nu} = e^{-ihu - i\bar{h}v} F_{\mu\nu}(\rho) \tag{39}$$

Now let's specialize to primary states $|\psi_{\mu\nu}\rangle$ which obey $L_1|\psi_{\mu\nu}\rangle = \bar{L}_1|\psi_{\mu\nu}\rangle = 0$. These conditions plus the gauge conditions give $h - \bar{h} = \pm 2$ and

$$F_{\mu\nu}(\rho) = f(\rho) \begin{pmatrix} 1 & \frac{h-\bar{h}}{2} & \frac{i}{\sinh(\rho)\cosh(\rho)} \\ \frac{h-\bar{h}}{2} & 1 & \frac{i(h-\bar{h})}{2\sinh(\rho)\cosh(\rho)} \\ \frac{i}{\sinh(\rho)\cosh(\rho)} & \frac{i(h-\bar{h})}{2\sinh(\rho)\cosh(\rho)} & -\frac{1}{\sinh^2(\rho)\cosh^2(\rho)} \end{pmatrix} \tag{40}$$

where

$$\partial_\rho f(\rho) + \frac{(h+\bar{h})\sinh^2\rho - 2\cosh^2\rho}{\sinh\rho\cosh\rho} f(\rho) = 0 \tag{41}$$

for the primary states. The solution is

$$f(\rho) = (\cosh\rho)^{-(h+\bar{h})} \sinh^2\rho \tag{42}$$

We see that $\sqrt{-\bar{g}}\psi^{\mu\nu *}\psi_{\mu\nu} = 4\sinh\rho\ (\cosh\rho)^{-3-2(h+\bar{h})}$. Also using $L^2|\psi_{\mu\nu}\rangle = -h(h-1)|\psi_{\mu\nu}\rangle$ for primaries, the primary weights $(h, \bar{h})$ obey:

$$[2h(h-1)+2\bar{h}(\bar{h}-1)-3-\mu^2\ell^2][h(h-1)+\bar{h}(\bar{h}-1)-2] = 0, \quad h-\bar{h} = \pm 2 \tag{43}$$

There are two branches of solutions. The first branch has $h(h-1) + \bar{h}(\bar{h}-1) - 2 = 0$, which gives:

$$h = \frac{3\pm 1}{2},\ \bar{h} = \frac{-1\pm 1}{2} \quad \text{or} \quad h = \frac{-1\pm 1}{2},\ \bar{h} = \frac{3\pm 1}{2} \tag{44}$$

These are the solutions that already appear in Einstein gravity. The solutions with the lower sign will blow up at the infinity, so we will only keep the upper ones corresponding to weights $(2,0)$ and $(0,2)$. We will refer to these as left and right-moving massless gravitons.

The second branch has $2h(h-1)+2\bar{h}(\bar{h}-1)-3-\mu^2\ell^2=0$, which gives:

$$h=\frac{3}{2}\pm\frac{\mu\ell}{2} \qquad \bar{h}=-\frac{1}{2}\pm\frac{\mu\ell}{2} \tag{45}$$

$$\text{or} \qquad h=-\frac{1}{2}\pm\frac{\mu\ell}{2} \qquad \bar{h}=\frac{3}{2}\pm\frac{\mu\ell}{2} \tag{46}$$

where (45) are the solutions of (23), and (46) are the solutions of (23) with μ replaced by $-\mu$. In the case of interest $\mu \geq 1/\ell$, only the solutions with the plus signs in (45) will not blow up at the infinity. Hence the relevant solutions corresponding to massive gravitons are

$$h=\frac{3}{2}+\frac{\mu\ell}{2}, \qquad \bar{h}=-\frac{1}{2}+\frac{\mu\ell}{2}\ . \tag{47}$$

Descendants are obtained by simply applying L_{-1} and $\bar{L}_{-1}$ on the primary $|\psi_{\mu\nu}\rangle$.

Note that for chiral gravity at $\mu\ell=1$, with $c_L=0,\ c_R=\frac{3l}{G}$, (47) becomes $h=2, \bar{h}=0$. Furthermore the wave function for the massive graviton becomes identical to that of the left-moving massless graviton. In fact, we can eliminate the $(2,0)$ modes with the residual gauge transformation

$$\epsilon_t = e^{-2iu}\frac{i\sinh^4(\rho)}{6\cosh^2(\rho)}\ , \tag{48}$$

$$\epsilon_\phi = e^{-2iu}\frac{-i\sinh^2(\rho)\left(2+\cosh^2(\rho)\right)}{6\cosh^2(\rho)}\ , \tag{49}$$

$$\epsilon_\rho = e^{-2iu}\frac{\sinh(\rho)\left(1+2\cosh^2(\rho)\right)}{6\cosh^3(\rho)}\ , \tag{50}$$

which satisfies

$$\bar{\nabla}_\mu\epsilon_\nu+\bar{\nabla}_\nu\epsilon_\mu+\psi^{(2,0)}_{\mu\nu} = 0\ . \tag{51}$$

Note that this gauge transformation does not vanish at the boundary, so whether or not the $(2,0)$ solution should be regarded as gauge equivalent to the vacuum depends on the so-far-unspecified boundary conditions.

4 Positivity of energy

4.1 BTZ black holes

The addition of the Chern-Simons term to the bulk action requires additional surface terms which in turn modify the definition of energy [5]. These surface terms are non-vanishing in general even for those solutions which satisfy the Einstein equation — namely BTZ black holes and massless gravitons. For Einstein metrics, the mass $M(\mu)$ and angular momentum $J(\mu)$ at general coupling μ are related to their values at $\mu = \infty$ by [27] [8],

$$\ell M(\mu) = \ell M(\infty) + \frac{J(\infty)}{\mu\ell} \tag{52}$$

$$J(\mu) = J(\infty) + \frac{M(\infty)}{\mu} \tag{53}$$

The bound $\ell M(\infty) \geq |J(\infty)|$ then implies positivity of energy for Einstein metrics when $\mu\ell > 1$. When $\mu\ell < 1$, the energy of the black holes can be negative (There is some discussion in this region in [29] [30], where signs of instability also appear). Note that for chiral gravity at $\mu\ell = 1$ we have

$$\ell M(\frac{1}{\ell}) = J(\frac{1}{\ell}). \tag{54}$$

This can be interpreted as the statement that all Einstein geometries are right-moving.

Now let's compute the entropy of the black hole, assuming the existence of a unitary dual CFT, following [7] [8]. The inner and outer horizons are at

$$r_{\pm} = \sqrt{2G\ell(\ell M(\infty) + J(\infty))} \pm \sqrt{2G\ell(\ell M(\infty) - J(\infty))} \tag{55}$$

and do not depend on μ. In terms of these the macroscopic formula for the entropy, including a contribution from the Chern-Simons term, is [24] [28] [29] [31].

$$S_{BH}(\mu) = \frac{\pi}{2G}(r_+ + \frac{1}{\mu\ell} r_-) \tag{56}$$

The left and right temperatures of the black hole are determined by periodicities and also do not depend on μ. They are

$$T_L = \frac{r_+ - r_-}{2\pi\ell^2}, \quad T_R = \frac{r_+ + r_-}{2\pi\ell^2} \tag{57}$$

In terms of these quantities, the microscopic Cardy formula for the entropy is

$$S_{BH}(\mu) = \frac{\pi^2 c_L(\mu) T_L \ell}{3} + \frac{\pi^2 c_R(\mu) T_R \ell}{3} \tag{58}$$

Using formula (13) for the central charges one readily finds that this agrees with the macroscopic result (56).

4.2 Gravitons

For $\mu\ell > 1$ the weights $(h, \bar{h})$ of the massive gravitons are positive. The energy of the massive gravitons is proportional to these weights but with a possible minus sign. In particular if the overall sign of the action is changed, so is the energy, but the equations of motion and hence the weights are unaffected. In [4] [5] [32], it was shown that with the sign taken herein but no cosmological constant massive gravitons have negative energy. In this section we redo this analysis for the case of negative cosmological constant, by constructing the Hamiltonian [33].

The fluctuation $h_{\mu\nu}$ can be decomposed as

$$h_{\mu\nu} \;=\; h^M_{\mu\nu} + h^L_{\mu\nu} + h^R_{\mu\nu} \tag{59}$$

where we use the subscript M to represent the $(\frac{3+\mu\ell}{2}, \frac{-1+\mu\ell}{2})$ primary and their descendants, L to represent the $(2,0)$ primary and their descendants, and R to represent the $(0,2)$ primary and their descendants. We will call them massive modes, left-moving modes and right-moving modes hereafter.

Up to total derivatives, the quadratic action of $h_{\mu\nu}$ is

$$\begin{aligned} S_2 &= -\frac{1}{32\pi G}\int d^3x\sqrt{-g}h^{\mu\nu}(\mathcal{G}^{(1)}_{\mu\nu} + \frac{1}{\mu}C^{(1)}_{\mu\nu}) \\ &= \frac{1}{64\pi G}\int d^3x\sqrt{-g}\{-\bar{\nabla}^\lambda h^{\mu\nu}\bar{\nabla}_\lambda h_{\mu\nu} + \frac{2}{\ell^2}h^{\mu\nu}h_{\mu\nu} - \frac{1}{\mu}\bar{\nabla}_\alpha h^{\mu\nu}\epsilon_\mu{}^{\alpha\beta}(\bar{\nabla}^2 + \frac{2}{\ell^2})h_{\beta\nu}\} \end{aligned} \tag{60}$$

The momentum conjugate to $h_{\mu\nu}$ is

$$\Pi^{(1)\mu\nu} = -\frac{\sqrt{-g}}{64\pi G}\left(\bar{\nabla}^0(2h^{\mu\nu} + \frac{1}{\mu}\epsilon^{\mu\alpha}{}_{\beta}\bar{\nabla}_\alpha h^{\beta\nu}) - \frac{1}{\mu}\epsilon_\beta{}^{0\mu}(\bar{\nabla}^2 + \frac{2}{\ell^2})h^{\beta\nu}\right) \quad (61)$$

Using the equations of motion we find,

$$\begin{aligned}
\Pi_M^{(1)\mu\nu} &= \frac{\sqrt{-g}}{64\pi G}(-\bar{\nabla}^0 h^{\mu\nu} + \frac{1}{\mu}(\mu^2 - \frac{1}{\ell^2})\epsilon_\beta{}^{0\mu}h_M^{\beta\nu}) && (62)\\
\Pi_L^{(1)\mu\nu} &= -\frac{\sqrt{-g}}{64\pi G}(2 - \frac{1}{\mu\ell})\bar{\nabla}^0 h_L^{\mu\nu} && (63)\\
\Pi_R^{(1)\mu\nu} &= -\frac{\sqrt{-g}}{64\pi G}(2 + \frac{1}{\mu\ell})\bar{\nabla}^0 h_R^{\mu\nu} && (64)
\end{aligned}$$

Because we have up to three time derivatives in the Lagrangian, using the Ostrogradsky method we should also introduce $K_{\mu\nu} \equiv \bar{\nabla}_0 h_{\mu\nu}$ as a canonical variable, whose conjugate momentum is

$$\Pi^{(2)\mu\nu} = \frac{-\sqrt{-g}g^{00}}{64\pi G\mu}\epsilon_\beta{}^{\lambda\mu}\bar{\nabla}_\lambda h^{\beta\nu} \quad (65)$$

again using equations of motion we have

$$\begin{aligned}
\Pi_M^{(2)\mu\nu} &= \frac{-\sqrt{-g}g^{00}}{64\pi G}h_M^{\mu\nu} && (66)\\
\Pi_R^{(2)\mu\nu} &= \frac{-\sqrt{-g}g^{00}}{64\pi G\mu\ell}h_L^{\mu\nu} && (67)\\
\Pi_L^{(2)\mu\nu} &= \frac{\sqrt{-g}g^{00}}{64\pi G\mu\ell}h_R^{\mu\nu} && (68)
\end{aligned}$$

The Hamiltonian is then

$$H = \int d^3x\big(\dot{h}_{\mu\nu}\Pi^{(1)\mu\nu} + \dot{K}_{i\mu}\Pi^{(2)i\mu} - S\big) \quad (69)$$

Specializing to linearized gravitons, and using their equations of motion, we then have the energies

$$\begin{aligned}
E_M &= \frac{1}{\mu}(\mu^2 - \frac{1}{\ell^2})\int d^3x\frac{\sqrt{-g}}{64\pi G}\epsilon_\beta{}^{0\mu}h_M^{\beta\nu}\dot{h}_{M\mu\nu} && (70)\\
E_L &= (-1 + \frac{1}{\mu\ell})\int d^3x\frac{\sqrt{-g}}{32\pi G}\bar{\nabla}^0 h_L^{\mu\nu}\dot{h}_{L\mu\nu} && (71)\\
E_R &= (-1 - \frac{1}{\mu\ell})\int d^3x\frac{\sqrt{-g}}{32\pi G}\bar{\nabla}^0 h_R^{\mu\nu}\dot{h}_{R\mu\nu} && (72)
\end{aligned}$$

All the integrals above are negative, as can be shown by plugging in the solutions (39), (40) and (42) for primaries and by using the $SL(2,\mathbb{R})$ algebra for descendants. For the massive mode, the energy is negative when $\mu\ell > 1$ (it becomes infinitely negative at $\mu = \infty$) and positive when $\mu\ell < 1$. The energy of left-moving modes is positive when $\mu\ell > 1$, and negative when $\mu\ell < 1$. The energy of right-moving modes is always positive. $\mu\ell = 1$ is a critical point, where the energy of both massive modes and the left-moving modes becomes zero. Note E_L and E_R are consistent with (52).

Recall that positivity of central charges requires that $\mu\ell \geq 1$, so the only possibility for avoiding negative energy is to take the chiral gravity theory with $\mu\ell = 1$. In that case, we see that massive and left-moving gravitons carry no energy, and might perhaps be regarded as pure gauge.

5 Conclusion and Discussion

In this note, we investigated the cosmological TMG, and found that the theory can be at most sensible at $\mu\ell = 1$. At this special point, the theory is completely chiral. In order to show that the chiral theory is classically sensible asymptotic boundary conditions which consistently eliminate the infinite degeneracy of zero-energy excitations must be specified. One must also prove a non-linear positive energy theorem. Renormalizability must be addressed to define the quantum theory.

At the classical level the chiral structure might enable an exact solution of the theory. Some exact non-Einstein solutions of TMG with arbitrary μ have been found in [34]-[40]. Should there turn out to also be a consistent quantum theory it would be interesting to find the chiral CFT dual. Towards this end the approach of [9] may prove useful.

Acknowledgements

We thank E. Witten and X. Yin for helpful discussions. The work is supported by DOE grant DE-FG02-91ER40654. WS is also supported by CSC. WS would like to thank the hospitality of the physics department of Harvard University.

APPENDIX: Energy-momentum pseudotensor

In this appendix we show that the energy defined from the energy momentum pseudotensor for massive gravitons can be negative. For simplicity we specialize to $\Lambda = 0$: for $\mu\ell \gg 1$ the Compton wavelength of the gravitons is much shorter than the AdS_3 radius so the latter can be locally ignored.

Let us first review how the energy-momentum pseudo tensor at quadratic order is defined without the Chern-Simons term or cosmological constant term. The full Einstein equation in the presence of matter is

$$G_{\mu\nu} = 16\pi G T^M_{\mu\nu} \tag{73}$$

Expanding the metric $g_{\mu\nu} = \eta_{\mu\nu} + h_{\mu\nu}$, we have through quadratic order

$$G^{(1)}_{\mu\nu} = -G^{(2)}_{\mu\nu} + 16\pi G T^M_{\mu\nu} \tag{74}$$

where $G^{(1)}_{\mu\nu}$ and $G^{(2)}_{\mu\nu}$ are terms linear and quadratic in $h_{\mu\nu}$. The energy momentum pseudo tensor defined as

$$t_{\mu\nu} = -\frac{1}{16\pi G} G^{(2)}_{\mu\nu}, \quad \text{with} \quad \partial^\mu t_{\mu\nu} = 0, \tag{75}$$

sources the Newtonian part of the gravitational potential in the same way that the matter stress tensor does. When adding the Chern-Simons term, the energy momentum pseudo tensor is similarly defined as

$$t_{\mu\nu} = -\frac{1}{16\pi G}(G^{(2)}_{\mu\nu} + \frac{1}{\mu} C^{(2)}_{\mu\nu}). \tag{76}$$

In flat background, the linearized equation of motion (23) becomes

$$\partial^2 h_{\mu\nu} + \frac{1}{\mu}\epsilon_\mu{}^{\alpha\beta}\partial_\alpha \partial^2 h_{\beta\nu} = 0 \tag{77}$$

under the harmonic plus traceless gauge. For a plane wave solution in the form of $h_{\mu\nu}(k) = \frac{1}{\sqrt{2k_0}} e^{-ik\cdot x} e_{\mu\nu}(k)$, the gauge conditions and the equations of motion are

$$\begin{aligned} k^\mu e_{\mu\nu}(k) &= 0,\ e^\mu{}_\mu(k) = 0 && (78) \\ k^2[e_{\mu\nu}(k) &- \frac{ik_\alpha}{\mu}\epsilon_\mu{}^{\alpha\beta} e_{\beta\nu}(k)] = 0. && (79) \end{aligned}$$

The $k^2 = 0$ solution is pure gauge. So $e_{\mu\nu}(k) - \frac{ik_\alpha}{\mu}\epsilon_\mu{}^{\alpha\beta} e_{\beta\nu}(k) = 0$, which implies $(k^2+\mu^2)e_{\mu\nu}(k) = 0$. When $k_\mu = (\mu, 0, 0)$, the positive energy solution is

$$h_{\mu\nu} = \frac{e^{-i\mu\tau}}{\sqrt{2\mu}} \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & i \\ 0 & i & -1 \end{pmatrix} \tag{80}$$

It is convenient to define $e_\mu(\vec{0}) = (0, 1, i)$, such that

$$\begin{aligned} e_{\mu\nu}(\vec{0}) &= e_\mu(\vec{0}) e_\nu(\vec{0}), && (81) \\ \text{and} \quad e_{\mu\nu}(\vec{k}) &= e_\mu(\vec{k}) e_\nu(\vec{k}), && (82) \end{aligned}$$

where $e_\mu(\vec{k})$ is obtained from $e_\mu(\vec{0})$ by a boost. The metric fluctuation can be expanded in Fourier modes,

$$h_{\mu\nu} = \int d\vec{k}^2 \frac{1}{\sqrt{2k_0}} \{\alpha(\vec{k}) e_{\mu\nu}(\vec{k}) e^{-ik\cdot x} + \alpha^\dagger(\vec{k}) e^*_{\mu\nu}(\vec{k}) e^{ik\cdot x}\} \tag{83}$$

To calculate the energy momentum pseudo tensor, we will need the following,

$$\begin{aligned} t_{\mu\nu} &= -\frac{1}{16\pi G}(\mathcal{G}^{(2)}_{\mu\nu} + \frac{1}{\mu} C^{(2)}_{\mu\nu}) && (84) \\ \mathcal{G}^{(2)}_{\mu\nu} &= R^{(2)}_{\mu\nu} - \frac{1}{2}\eta_{\mu\nu} R^{(2)} && (85) \\ R^{(2)}_{\mu\nu} &= \frac{1}{2} h^{\rho\sigma}\partial_\mu\partial_\nu h_{\rho\sigma} - h^{\rho\sigma}\partial_\rho\partial_{(\mu} h_{\nu)\sigma} + \frac{1}{4}(\partial_\mu h_{\rho\sigma})\partial_\nu h^{\rho\sigma} && (86) \\ &+ (\partial^\sigma h^\rho{}_\nu)\partial_{[\sigma} h_{\rho]\mu} + \frac{1}{2}\partial_\sigma(h^{\rho\sigma}\partial_\rho h_{\mu\nu}) && (87) \\ C^{(2)}_{\mu\nu} &= \epsilon_\mu{}^{\alpha\beta}\partial_\alpha(R^{(2)}_{\beta\nu} - \frac{1}{4}\eta_{\beta\nu} R^{(2)}) + h_{\mu\lambda}\epsilon^{\lambda\alpha\beta}\partial_\alpha R^{(1)}_{\beta\nu} - \epsilon_\mu{}^{\alpha\beta}\Gamma^{\lambda(1)}_{\alpha\nu} R^{(1)}_{\beta\lambda} && (88) \\ \Gamma^{\lambda(1)}_{\nu\alpha} &= \frac{1}{2}\eta^{\lambda\rho}(\partial_\alpha h_{\rho\nu} + \partial_\nu h_{\rho\alpha} - \partial_\rho h_{\alpha\nu}) && (89) \\ R^{(1)}_{\mu\nu} &= -\frac{1}{2}\partial^2 h_{\mu\nu} && (90) \end{aligned}$$

Using the equation of motion (32), it simplifies to

$$G^{(2)}_{\mu\nu} = -\frac{1}{4}(\partial_\mu h_{\rho\sigma})\partial_\nu h^{\rho\sigma} - \frac{\mu^2}{2}h^\rho{}_\mu h_{\rho\nu} - \frac{\mu^2}{8}\eta_{\mu\nu}h^{\rho\sigma}h_{\rho\sigma}, \tag{91}$$

$$C^{(2)}_{\mu\nu} = \frac{\mu^2}{4}(3\mu h_{\rho\mu}h^\rho{}_\nu + \epsilon_\mu{}^{\alpha\beta}h^\lambda{}_{\alpha,\nu}h_{\beta\lambda}) \tag{92}$$

up to total derivatives. So the energy is

$$E = \int d^2\vec{x}\, t_{00} \tag{93}$$

$$= \int d^2\vec{k}\frac{n(k)}{16\pi G k_0}\{(k_0^2 - \frac{\mu^2}{2} - k_0\mu) - \frac{1}{2}\mu^2 e_0(\vec{k})^* e_0(\vec{k})\}. \tag{94}$$

For vanishing spatial momentum $n(\vec{k}) \propto \delta^2(\vec{k})$, the energy is just $-\mu$ times a positive normalization factor.

References

[1] M. Banados, C. Teitelboim and J. Zanelli, "The Black hole in three-dimensional space-time," Phys. Rev. Lett. **69**, 1849 (1992) [arXiv:hep-th/9204099].

[2] E. Witten, "(2+1)-Dimensional Gravity as an Exactly Soluble System," Nucl. Phys. B **311**, 46 (1988).

[3] J. Fjelstad and S. Hwang, "Sectors of solutions in three-dimensional gravity and black holes," Nucl. Phys. B **628**, 331 (2002) [arXiv:hep-th/0110235].

[4] S. Deser, R. Jackiw and S. Templeton, Annals Phys. **140**, 372 (1982) [Erratum-ibid. **185**, 406.1988 APNYA,281,409 (1988 APNYA,281,409-449.2000)]

[5] S. Deser, R. Jackiw and S. Templeton, "Three-Dimensional Massive Gauge Theories," Phys. Rev. Lett. **48**, 975 (1982).

[6] A. Strominger, "Black hole entropy from near-horizon microstates," JHEP **9802**, 009 (1998) [arXiv:hep-th/9712251].

[7] P. Kraus and F. Larsen, "Microscopic black hole entropy in theories with higher derivatives," JHEP **0509**, 034 (2005) [arXiv:hep-th/0506176].

[8] P. Kraus and F. Larsen, "Holographic gravitational anomalies," JHEP **0601**, 022 (2006) [arXiv:hep-th/0508218].

[9] E. Witten, "Three-Dimensional Gravity Revisited," arXiv:0706.3359 [hep-th].

[10] J. Manschot, "AdS_3 Partition Functions Reconstructed," JHEP **0710**, 103 (2007) [arXiv:0707.1159 [hep-th]].

[11] D. Gaiotto and X. Yin, "Genus Two Partition Functions of Extremal Conformal Field Theories," JHEP **0708**, 029 (2007) [arXiv:0707.3437 [hep-th]].

[12] M. R. Gaberdiel, "Constraints on extremal self-dual CFTs," JHEP **0711**, 087 (2007) [arXiv:0707.4073 [hep-th]].

[13] S. D. Avramis, A. Kehagias and C. Mattheopoulou, "Three-dimensional AdS gravity and extremal CFTs at c=8m," JHEP **0711**, 022 (2007) [arXiv:0708.3386 [hep-th]].

[14] X. Yin, "Partition Functions of Three-Dimensional Pure Gravity," arXiv:0710.2129 [hep-th].

[15] X. Yin, "On Non-handlebody Instantons in 3D Gravity," arXiv:0711.2803 [hep-th].

[16] A. Maloney and E. Witten, "Quantum Gravity Partition Functions in Three Dimensions," arXiv:0712.0155 [hep-th].

[17] S. Deser and B. Tekin, "Massive, topologically massive, models," Class. Quant. Grav. **19**, L97 (2002) [arXiv:hep-th/0203273].

[18] J. Maldacena and A. Strominger "Ads3 Black Holes and a Stringy Exclusion Principle" JHEP **9812**, 005 (1998) [arxiv:hep-th/9804085].

[19] J. D. Brown and M. Henneaux, "Central charges in the canonical realization of asymptotic symmetries: an example from three-dimensional gravity," Commun. Math. Phys. 104, 207(1986).

[20] M. Henningson and K. Skenderis, "The holographic Weyl anomaly," JHEP **9807**, 023 (1998) [arXiv:hep-th/9806087].

[21] S. de Haro, S. N. Solodukhin and K. Skenderis, "Holographic reconstruction of spacetime and renormalization in the AdS/CFT correspondence," Commun. Math. Phys. **217**, 595 (2001) [arXiv:hep-th/0002230].

[22] V. Balasubramanian and P. Kraus, "A stress tensor for anti-de Sitter gravity," Commun. Math. Phys. **208**, 413 (1999) [arXiv:hep-th/9902121].

[23] P. Kraus, "Lectures on black holes and the AdS(3)/CFT(2) correspondence," arXiv:hep-th/0609074.

[24] S. N. Solodukhin, "Holography with gravitational Chern-Simons," Phys. Rev. D **74**, 024015 (2006) [arXiv:hep-th/0509148].

[25] S. Deser and B. Tekin, "Energy in topologically massive gravity," Class. Quant. Grav. **20** (2003) L259 [arXiv:gr-qc/0307073].

[26] S. Olmez, O. Sarioglu and B. Tekin, "Mass and angular momentum of asymptotically AdS or flat solutions in the topologically massive gravity," Class. Quant. Grav. **22** (2005) 4355 [arXiv:gr-qc/0507003].

[27] K. A. Moussa, G. Clement and C. Leygnac, "The black holes of topologically massive gravity," Class. Quant. Grav. **20**, L277 (2003) [arXiv:gr-qc/0303042].

[28] B. Sahoo and A. Sen, "BTZ black hole with Chern-Simons and higher derivative terms," JHEP **0607**, 008 (2006) [arXiv:hep-th/0601228].

[29] M. I. Park, "BTZ black hole with gravitational Chern-Simons: Thermodynamics and statistical entropy," Phys. Rev. D **77**, 026011 (2008) arXiv:hep-th/0608165.

[30] M. I. Park, Phys. Lett. B **647**, 472 (2007) [arXiv:hep-th/0602114]; [arXiv:hep-th/0610140]; [arXiv:hep-th/0609027]; [arXiv:hep-th/0611048].

[31] Y. Tachikawa, "Black hole entropy in the presence of Chern-Simons terms," Class. Quant. Grav. **24**, 737 (2007) [arXiv:hep-th/0611141].

[32] S. Deser and J. H. Kay, "Topologically Massive Supergravity," Phys. Lett. B **120**, 97 (1983).

[33] I. L. Buchbinder, S. L. Lyahovich and V. A. Krychtin, "Canonical Quantization Of Topologically Massive Gravity," Class. Quant. Grav. **10**, 2083 (1993).

[34] R. Percacci, P. Sodano and I. Vuorio, "Topologically massive planer universes with constant twist," Annals Phys. **176**, 344 (1987).

[35] G. S. Hall, T. Morgan and Z. Perjes, "Three-dimensional space-times,"

[36] Y. Nutku, "Exact solutions of topologically massive gravity with a cosmological constant," Class. Quant. Grav. **10**, 2657 (1993).

[37] A. N. Aliev and Y. Nutku, "A theorem on topologically massive gravity," Class. Quant. Grav. **13**, L29 (1996) [arXiv:gr-qc/9812089].

[38] A. N. Aliev and Y. Nutku, "Spinor formulation of topologically massive gravity," arXiv:gr-qc/9812090.

[39] T. Dereli and O. Sarioglu, "Topologically massive gravity and black holes in three dimensions," arXiv:gr-qc/0009082.

[40] A. Bouchareb and G. Clement, "Black hole mass and angular momentum in topologically massive gravity," Class. Quant. Grav. **24**, 5581 (2007) [arXiv:0706.0263 [gr-qc]].

arXiv:0903.4573v1 [hep-th] 26 Mar 2009

Chiral Gravity, Log Gravity and Extremal CFT

Alexander Maloney[†], Wei Song[‡§] and Andrew Strominger[§]

[†] *Physics Department, McGill University, Montreal, QC H3A 2T8, Canada*
[‡] *Key Laboratory of Frontiers in Theoretical Physics,*
Institute of Theoretical Physics, Chinese Academy of Sciences, Beijing, 100190, China
[§] *Center for the Fundamental Laws of Nature*
Jefferson Physical Laboratory, Harvard University, Cambridge, MA 02138, USA

Abstract

We show that the linearization of all exact solutions of classical chiral gravity around the AdS_3 vacuum have positive energy. Non-chiral and negative-energy solutions of the linearized equations are infrared divergent at second order, and so are removed from the spectrum. In other words, chirality is confined and the equations of motion have linearization instabilities. We prove that the only stationary, axially symmetric solutions of chiral gravity are BTZ black holes, which have positive energy. It is further shown that classical log gravity – the theory with logarithmically relaxed boundary conditions – has finite asymptotic symmetry generators but is not chiral and hence may be dual at the quantum level to a logarithmic CFT. Moreover we show that log gravity contains chiral gravity within it as a decoupled charge superselection sector. We formally evaluate the Euclidean sum over geometries of chiral gravity and show that it gives precisely the holomorphic extremal CFT partition function. The modular invariance and integrality of the expansion coefficients of this partition function are consistent with the existence of an exact quantum theory of chiral gravity. We argue that the problem of quantizing chiral gravity is the holographic dual of the problem of constructing an extremal CFT, while quantizing log gravity is dual to the problem of constructing a logarithmic extremal CFT.

Contents

1 Introduction 1

2 Preliminaries 5

3 Symmetry generators 7

4 Classical perturbation theory 9
- 4.1 Chirality confinement . . . 10
- 4.2 The CDWW modes . . . 13
- 4.3 The GKP mode . . . 15
- 4.4 Global modes . . . 16

5 A Birkhoff-like theorem 17
- 5.1 Stationarity and axial symmetry . . . 17
- 5.2 Boundary conditions . . . 19
- 5.3 Solutions . . . 20

6 Log gravity 25
- 6.1 Boundary conditions and non-chirality . . . 25
- 6.2 Decoupling the chiral gravity superselection sector . . . 27

7 Quantum partition function 29
- 7.1 Classical saddle points . . . 30
- 7.2 Sum over geometries . . . 33
- 7.3 Physical interpretation . . . 36
- 7.4 What could go wrong . . . 38

8 Chiral gravity, log gravity, extremal CFT and log extremal CFT 38
- 8.1 Nothing makes sense . . . 38
- 8.2 Chiral gravity = extremal CFT . . . 39
- 8.3 Chiral gravity $\in$ log gravity=log extremal CFT . . . 39

References 41

1 Introduction

A consistent, non-trivial theory of pure gravity in three dimensions – classical or quantum – with a stable vacuum would undoubtedly provide invaluable insights into the many complexities of gravity in our four-dimensional world. Unfortunately, pure 3D Einstein gravity is locally trivial classically, while its quantum status remains

unclear despite decades of investigations. Recently, an exceptional and clearly non-trivial 3D theory termed "chiral gravity" was discovered [1]. This theory is a special case of topologically massive gravity [2, 3] at a particular value of the couplings, and is defined with asymptotically AdS_3 boundary conditions, in the sense of Fefferman-Graham-Brown-Henneaux [4, 5].[1] It was conjectured in [1] that at the classical level

- Chiral gravity is chiral, in the sense that the asymptotic symmetry group is generated by a single copy of the Virasoro algebra, [2]
- Solutions of chiral gravity have positive energy.

Some supporting evidence was given [1]. Should both conjectures turn out to be true, chiral gravity, in its quantum version, would prove an extremely interesting gedanken laboratory for the study of quantum gravity.

The chirality and positivity conjectures generated some controversy. Shortly after [1], interesting new solutions to the linearized equations which are not global energy eigenmodes and have a variety of asymptotic behaviors were discovered. These solutions are non-chiral and/or negative-energy and were argued to provide counterexamples to one or both of the classical conjectures [6, 8] – see also [10, 11, 12, 13, 14]. Subsequently the chirality conjecture was proven [15] and the claims that these modes provide counterexamples to this conjecture were revised or withdrawn [7, 9]. A proof of the chirality conjecture in a different formalism appeared in [16]. Nevertheless, claims that the proposed counterexamples disprove the positivity conjecture remain in the literature. However the modes exhibited in [6, 8] explicitly violate the chirality conjecture as well as the positivity conjecture. So if these modes are truly present in the linearization of the exact spectrum, they are fatally at odds not only with positivity but with the chirality proofs of [15, 16]. In short, the literature contains contradictory claims. For related work, see [17, 18, 19, 20, 21, 22, 23, 24, 25].

In this paper, we reconcile all these computations and hope to thereby resolve the controversy. In the process, a perturbative version of the positivity conjecture will be established to first order in the deviation around AdS_3. The alleged counterexamples do not disprove positivity for exactly the same reason that they do not disprove chirality: the equations have a linearization instability. At second order in perturbation theory, explicit computation reveals that the metric perturbation develops an

[1] Chiral gravity differs in this respect from log gravity which has the same action but logarithmically weaker boundary conditions.

[2] The quantum version of this conjecture is that physical states form representations of a single Virasoro algebra.

infrared singularity, growing logarithmically with the radius at infinity. Hence these solutions of the linearized equations are *not* the linear approximation to any exact solution of the theory with Brown-Henneaux boundary conditions. In other words, chirality is confined and chiral gravity has linearization instabilities.[3] This divergence was bound to appear because otherwise there would be a discrepancy between the surface integral expression for the energy (linear in the second order perturbation) and the bulk expression (quadratic in the first order perturbation). The first of these is manifestly chiral for asymptotically AdS_3 solutions, while the second gives a non-chiral answer. A key ingredient in reconciling herein the various computations is the discovery (independently made in [29]) of previously neglected terms in the boundary expressions for the Virasoro charges. The omission of these terms has led to some contradictory statements in the literature.

An (imperfect) analogy can be found in QCD. In the linearized approximation, the theory contains free quarks. But there is an infrared divergence in the back reaction caused by the quark and the exact finite energy spectrum contains only color singlets. A free quark is not a valid linearized approximation to any finite energy QCD state. Of course, if the boundary conditions are relaxed to allow flux tubes at infinity there are single-quark solutions. We will see below that an analogous relaxation of the boundary conditions for chiral gravity to those of log gravity allows for non-chiral excitations with finite charges.

The analogy here is imperfect in that color confinement in QCD is a difficult non-perturbative problem. In contrast, confinement in chiral gravity can be seen explicitly in second order perturbation theory. Moreover, in QCD color confinement gives one global constraint, while in chiral gravity there are an infinite number of constraints arising from the infinity of conserved (left) Virasoro charges, all of which must vanish. This is exactly what is required to eliminate an entire chiral half of the spectrum, and reconcile the chiral nature of the theory with the non-chiral "bulk" degree of freedom found in the local analyses of [6, 10, 11, 12, 13, 16, 8, 30]. Rather, we will see below these local analyses apply to log gravity.

The miraculous escape of chiral gravity from the alleged perturbative instabilities leads one to hope that there is an exact positive energy theorem for the theory. The proof of such a theorem at the non-perturbative level remains an outstanding challenge. We take one step in this direction by proving a Birkhoff-like theorem: all stationary, axially symmetric solutions of chiral gravity are BTZ black holes. The

[3] Similar linearization instabilities have occurred in a number of contexts in general relativity, see e.g. [26, 27, 28].

difficulty we encounter in what would seem a straightforward exercise illustrates the complexity of the full nonlinear equations. It is interesting to note that all known solutions of chiral gravity are also solutions of the Einstein equation. This may be the case for all solutions, although we will not attempt to demonstrate this here. One might also attempt to prove a version of cosmic censorship for chiral gravity.

Armed with knowledge the perturbative spectrum, we then move on to an analysis of the quantum problem. We apply the standard methodology of Euclidean quantum gravity to compute the torus partition function as a function of the modular parameter τ. Euclidean quantum gravity is, for a variety of reasons, a notoriously treacherous subject and the present application cannot be regarded as completely rigorous. Nevertheless the results are highly encouraging. We show that all real saddle points solve the Einstein equation, and can be classified. Moreover, at the chiral value of the coupling constants the Euclidean action is holomorphic. Following [31], we perform the sum over saddle points including all perturbative corrections, formally obtaining the exact answer for the partition function. The result is simply the "chiral part" of the extremal partition function conjectured by Witten [32] to be dual to 3D Einstein gravity. It is invariant under modular transformations and has an expansion in $q = e^{2\pi i\tau}$ with integer coefficients, as required for a consistent quantum mechanical interpretation as a Hilbert space trace. The spectrum reproduces the entropy of the BTZ black hole, including both the Bekenstein-Hawking piece and an infinite series of corrections. Although it is not known whether a CFT exists which realizes this spectrum, the encouraging outcome of this computation might be regarded as evidence both for the existence of quantum chiral gravity as well as for the existence of such CFTs. In any case the interesting problems of understanding quantum chiral gravity and extremal CFTs are clearly closely linked.

We also consider the theory of log gravity introduced in [10, 33]. This theory has the same action as chiral gravity, but the boundary conditions are weakened to allow metric fluctuations which grow logarithmically with the proper radius. Log gravity contains a rich and interesting class of solutions [34, 13, 21, 35] which are excluded in chiral gravity. In particular, the linearization of the exact spectrum includes the non-chiral modes of [6, 8, 10], which appear in indecomposable Virasoro representations. The relaxed boundary conditions also lead to zero-norm states, violations of unitary and violations of positivity. Interestingly, these violations resemble those found in logarithmic CFTs, suggesting that log gravity is dual to a logarithmic CFT [10]. We show here that the log gravity boundary conditions lead to finite expressions for the asymptotic symmetry generators. However, contrary to [33], the generators

are *not* chiral. This is consistent with the conjecture that log gravity is dual to a logarithmic CFT, as logarithmic conformal field theories cannot be chiral. We also show that log gravity contains within it chiral gravity as the superselection sector with vanishing left Virasoro charges. Thus although log gravity itself is not unitary, it has a potentially unitary "physical subspace". We speculate herein that log gravity may be dual to an "extremal" logarithmic CFT whose partition function coincides with Witten's extremal partition function.

This paper is organized as follows. Section 2 contains basic formulae and conventions. In section 3 we give the new expression for the asymptotic symmetry generators. In section 4 we work out the perturbation expansion around AdS_3 to second order, and show that the non-chiral negative energy solutions to the linearized equations are not the linearization of exact solutions. In section 5 we study the spectrum at the non-linear level, and prove a Birkhoff-like theorem for stationary, axially symmetric solutions. In section 6 we study log gravity, show that the asymptotic symmetry group has finite generators and discuss the problem of constructing a symplectic form as required for a canonical formulation. We show that although log gravity is non-chiral, it contains chiral gravity as a superselection sector. In section 7 we evaluate the Euclidean partition function and show that it gives the modular invariant extremal partition function. Finally section 8 concludes with a discussion of and speculations on the fascinating relation between chiral gravity, log gravity, extremal CFT and extremal logarithmic CFT.

As this work was nearing completion, the eprint [29] appeared with results which overlap with sections 3 and 6.1. All points in common are in precise agreement.

2 Preliminaries

In this section we record some pertinent formulae and establish notation. Chiral gravity is a special case of topologically massive gravity (TMG) [2, 3] with a negative cosmological constant. TMG is described by the action

$$I_{TMG} = \frac{1}{16\pi G}\left[\int d^3x\sqrt{-g}(R+2/\ell^2) + \frac{1}{\mu}I_{CS}\right] \tag{2.1}$$

where I_{CS} is the gravitational Chern-Simons action

$$I_{CS} = \frac{1}{2}\int_{\mathcal{M}} d^3x\sqrt{-g}\varepsilon^{\lambda\mu\nu}\Gamma^r_{\lambda\sigma}\left(\partial_\mu\Gamma^\sigma_{r\nu} + \frac{2}{3}\Gamma^\sigma_{\mu\tau}\Gamma^\tau_{\nu r}\right) \tag{2.2}$$

and G has the conventional positive sign. The equation of motion in TMG is

$$E_{\mu\nu} \equiv \mathcal{G}_{\mu\nu} + \frac{1}{\mu} C_{\mu\nu} = 0, \tag{2.3}$$

where we have defined

$$C_{\mu\nu} \equiv \epsilon^{\alpha\beta}{}_{(\mu} \mathcal{G}_{\nu)\beta;\alpha}, \qquad \mathcal{G}_{\mu\nu} \equiv G_{\mu\nu} - \frac{1}{\ell^2} g_{\mu\nu}. \tag{2.4}$$

These equations have the vacuum solution

$$\begin{aligned} ds^2 &= \ell^2 \left(-\cosh^2 \rho d\tau^2 + \sinh^2 \rho d\phi^2 + d\rho^2 \right) \\ &= \frac{\ell^2}{4} \left(-2\cosh 2\rho d\tau^+ d\tau^- - d\tau^{+2} - d\tau^{-2} + 4 d\rho^2 \right), \\ &\qquad \tau^\pm = \tau \pm \phi. \end{aligned} \tag{2.5}$$

Chiral gravity [1] is defined by taking $\mu\ell \to 1$ while keeping the standard Brown-Henneaux [5] asymptotically AdS_3 boundary conditions. These require that fluctuations $h_{\mu\nu}$ of the metric about (2.5) fall off at the boundary according to

$$\begin{pmatrix} h_{++} = \mathcal{O}(1) & h_{+-} = \mathcal{O}(1) & h_{+\rho} = \mathcal{O}(e^{-2\rho}) \\ h_{-+} = h_{+-} & h_{--} = \mathcal{O}(1) & h_{-\rho} = \mathcal{O}(e^{-2\rho}) \\ h_{\rho+} = h_{+\rho} & h_{\rho-} = h_{-\rho} & h_{\rho\rho} = \mathcal{O}(e^{-2\rho}) \end{pmatrix} \tag{2.6}$$

The consistency of these boundary conditions for all values of μ was demonstrated in [36]. The most general diffeomorphism which preserves (2.6) is of the form

$$\begin{aligned} \zeta &= \zeta^+ \partial_+ + \zeta^- \partial_- + \zeta^\rho \partial_\rho \\ &= \left[\epsilon^+(\tau^+) + 2e^{-2\rho} \partial_-^2 \epsilon^-(\tau^-) + \mathcal{O}(e^{-4\rho}) \right] \partial_+ \\ &+ \left[\epsilon^-(\tau^-) + 2e^{-2\rho} \partial_+^2 \epsilon^+(\tau^+) + \mathcal{O}(e^{-4\rho}) \right] \partial_- \\ &- \frac{1}{2} \left[\partial_+ \epsilon^+(\tau^+) + \partial_- \epsilon^-(\tau^-) + \mathcal{O}(e^{-2\rho}) \right] \partial_\rho. \end{aligned} \tag{2.7}$$

These are parameterized by a left moving function $\epsilon^-(\tau^-)$ and a right moving function $\epsilon^+(\tau^+)$. We denote diffeomorphisms depending only on ϵ^- by ξ^L and those depending only on ϵ^+ as ξ^R. The subleading terms all correspond to trivial diffeomorphisms; their generators have no surface term and hence vanish when the constraints are imposed. The asymptotic symmetry group (ASG) is defined as the general boundary-condition-preserving diffeomorphism (2.7) modulo the trivial diffeomorphisms. For

generic μ the ASG is generated by two copies of the Virasoro algebra, which may be taken to be

$$\xi_n^L = \xi(\epsilon^- = e^{in(\tau-\phi)}, \epsilon^+ = 0) \qquad \xi_n^R = \xi(\epsilon^- = 0, \epsilon^+ = e^{in(\tau+\phi)}). \tag{2.8}$$

These of course have a global $SL(2,\mathbb{R})_L \times SL(2,\mathbb{R})_R$ subgroup which generates the AdS_3 isometries. At the chiral point $\mu\ell = 1$ the left moving generators parameterized by $\epsilon^-(x^-)$ also become trivial [15]. Hence there is an enhancement of the trivial symmetry group and the ASG is generated by a single chiral Virasoro algebra.

3 Symmetry generators

In this section we present a refined expression for the symmetry generators which corrects expressions appearing in some of the literature.[4] The corrections are relevant only when the Brown-Henneaux boundary conditions are violated. This corrected expression is essential for demonstrating the general equality of the bulk and boundary expressions for the energy, as well as for the discussion of log gravity in section 6.

Our expression follows from the covariant formalism [37, 38], which is based on [39] and has been developed in great detail for a wide variety of applications in [40].[5] Let $E^{(1)}_{\mu\nu}(h)$ denote the linearization of the equation of motion (2.3) about AdS_3 metric $\bar{g}$ with respect to a small perturbation h near the boundary. One may then define the one-form

$$\mathcal{K}(\xi, h) \equiv \xi^\mu E^{(1)}_{\mu\nu}(h) dx^\nu \tag{3.1}$$

It is shown in [42] that when ξ is a Killing vector[6]

$$\mathcal{K}(\xi, h) = *d * \mathcal{F}(\xi, h). \tag{3.2}$$

Here $\mathcal{F}$ is a two form "superpotential," which is written out explicitly in [44]. It was further shown that the conserved charges associated to the ASG are then given by

[4] The expressions herein were independently found using a different formalism in [29].

[5] Some recent discussions of TMG have used the Brown-York formalism [41], which was initially developed for diffeomorphisms which – unlike those in (3.5) – do not have components normal to the boundary. For Brown-Henneaux boundary conditions this subtlety turns out to be irrelevant. It is, however, relevant when violations of the Brown-Henneaux boundary conditions are considered. While the Brown-York formalism could likely be generalized to this case, the covariant formalism is more highly developed and hence more convenient.

[6] This formalism was discussed for general backgrounds in [43], and further generalized in to the case where ξ is not an asymptotic Killing vector [44]. In this case an additional term appears on the left hand side of (3.2).

the boundary integral

$$Q_{boundary}(\xi) = -\frac{1}{16\pi G}\oint_{\partial\Sigma} *\mathcal{F}. \tag{3.3}$$

Here $\partial\Sigma$ is the boundary of a spacelike surface Σ. Integrating by parts gives the bulk expression

$$Q_{bulk}(\xi) = -\frac{1}{16\pi G}\int_{\Sigma} *K = Q_{boundary}(\xi) \tag{3.4}$$

In this formula K can be taken to be any smooth extension of the boundary one-form (3.1) into the interior. In the coordinates (2.5), we shall see in the next section that

$$\begin{aligned} Q_{boundary}(\xi) = \frac{1}{32\pi\ell G}\oint_{\partial\Sigma} d\phi \Bigg[\epsilon^- \left(-2\partial_\rho^2 h_{--} + 4\partial_\rho h_{--} + 2\partial_\rho h_{-+} - 4h_{-+} + \frac{e^{2\rho}}{4}h_{\rho\rho}\right) \\ + \epsilon^+ \left(8h_{++} - 8\partial_\rho h_{++} + 2\partial_\rho^2 h_{++} + 2\partial_\rho h_{-+} - 4h_{-+} + \frac{e^{2\rho}}{4}h_{\rho\rho}\right)\Bigg] \end{aligned} \tag{3.5}$$

In the above expression (3.5), we have only assumed that h falls off fast enough for Q to be finite, but have not used the Brown-Henneaux boundary condition (2.6). Asymptotically, the $\rho\rho$ component of the linearized equation of motion gives

$$2\partial_\rho h_{-+} - 4h_{-+} + \frac{e^{2\rho}}{4}h_{\rho\rho} = 0 \tag{3.6}$$

Condition (3.6) is an asymptotic constraint, as it involves only the fields and not their time derivatives and hence weakly vanishes in the Dirac bracket formalism. See [45] for a similar discussion in Gödel spacetime. Using the stricter boundary conditions (2.6), and imposing the asymptotic constraints, the expression becomes simply

$$Q_{boundary}(\xi) = \frac{1}{4\pi\ell G}\oint_{\partial\Sigma} d\phi \epsilon^+ h_{++}. \tag{3.7}$$

These charges can be decomposed into left and right charges Q^L and Q^R generating left and right diffeomorphisms $\xi^L(\epsilon^-)$ and $\xi^R(\epsilon^+)$. Note that ϵ^- does not appear in (3.7), so for Brown-Henneaux boundary conditions the $\xi^L(\epsilon^-)$ are trivial and the left charges vanish. This implies that the theory is chiral [15]:

$$Q^L \equiv Q(\xi^L) = 0. \tag{3.8}$$

Hence the name chiral gravity.

In the following we will study violations of the asymptotic boundary conditions where the extra terms in $Q_{boundary}$ will contribute. In particular, we will encounter

situations in which the $\partial_\rho h_{--}$ term above does not vanish and $Q^L \neq 0$. In this case the left moving charges can be written in a simple gauge invariant form

$$Q^L_{boundary} = \frac{1}{8\pi G}\oint_{\partial\Sigma} \xi^{L\mu}(\mathcal{G}^{(1)}_{\mu\nu} - \frac{g_{\mu\nu}}{2}\mathcal{G}^{(1)})dx^\nu, \tag{3.9}$$

We see that the left charges are nonzero only if the curvature perturbation does not vanish on the boundary.

4 Classical perturbation theory

In this section we will work out the weak field perturbation expansion of the equations of motion to second order. We start by expanding the metric around the AdS_3 background as

$$g_{\mu\nu} = \bar{g}_{\mu\nu} + h_{\mu\nu} = \bar{g}_{\mu\nu} + h^{(1)}_{\mu\nu} + h^{(2)}_{\mu\nu} + \cdots \tag{4.1}$$

The expansion parameter here is the magnitude of the first order fluctuation $h^{(1)}$. Inserting this into the full equation of motion

$$\mathcal{G}_{\mu\nu} + \frac{1}{\mu}C_{\mu\nu} = 0 \tag{4.2}$$

and expanding to first order in the perturbation we see that $h^{(1)}$ must satisfy

$$E^{(1)}_{\mu\nu}(h^{(1)}) \equiv \mathcal{G}^{(1)}_{\mu\nu}(h^{(1)}) + \frac{1}{\mu}C^{(1)}_{\mu\nu}(h^{(1)}) = 0 \tag{4.3}$$

In this and the following equations indices are raised and lowered using the background metric. The second order perturbation $h^{(2)}$ is found by expanding (4.2) to second order

$$E^{(1)}_{\mu\nu}(h^{(2)}) = -E^{(2)}_{\mu\nu}(h^{(1)}) \tag{4.4}$$

Explicit computation gives the left hand side of (4.4)

$$E^{(1)}_{\mu\nu} = \mathcal{G}^{(1)}_{\mu\nu} + \frac{1}{2\mu}(\epsilon_\mu{}^{\alpha\beta}\nabla_\alpha\mathcal{G}^{(1)}_{\nu\beta} + \epsilon_\nu{}^{\alpha\beta}\nabla_\alpha\mathcal{G}^{(1)}_{\mu\beta}) \tag{4.5}$$

$$\mathcal{G}^{(1)}_{\mu\nu} = R^{(1)}_{\mu\nu} + \frac{2}{\ell^2}h_{\mu\nu} - \frac{1}{2}g_{\mu\nu}(R^{(1)} + \frac{2}{\ell^2}h) \tag{4.6}$$

where

$$R^{(1)}_{\mu\nu} = \frac{1}{2}(-\nabla^2 h_{\mu\nu} - \nabla_\mu\nabla_\nu h + \nabla^\lambda\nabla_\nu h_{\mu\lambda} + \nabla^\lambda\nabla_\mu h_{\nu\lambda}) \tag{4.7}$$

$$\Gamma^{(1)\lambda}_{\mu\nu} = \frac{1}{2}\left[\nabla_\mu h^\lambda_\nu + \nabla_\nu h^\lambda_\mu - \nabla^\lambda h_{\mu\nu}\right]. \tag{4.8}$$

The right hand side of (4.4) is

$$\begin{aligned} E^{(2)}_{\mu\nu} &= \mathcal{G}^{(2)}_{\mu\nu} + \frac{1}{2\mu}[(\epsilon_\mu{}^{\alpha\beta}\nabla_\alpha \mathcal{G}^{(2)}_{\beta\nu} + h_{\mu\lambda}\epsilon^{\lambda\alpha\beta}\nabla_\alpha \mathcal{G}^{(1)}_{\beta\nu} - \frac{h}{2}\epsilon_\mu{}^{\alpha\beta}\nabla_\alpha \mathcal{G}^{(1)}_{\beta\nu} \\ &\quad -\epsilon_\mu{}^{\alpha\beta}\Gamma^{(1)\lambda}_{\nu\alpha}\mathcal{G}^{(1)}_{\beta\lambda}) + (\mu \leftrightarrow \nu)] \end{aligned} \tag{4.9}$$

$$\mathcal{G}^{(2)}_{\mu\nu} = R^{(2)}_{\mu\nu} - \frac{g_{\mu\nu}}{2}(R^{(2)} - h^{\lambda\sigma}R^{(1)}_{\lambda\sigma} + h^{\lambda\alpha}h^\sigma_\alpha R_{\lambda\sigma}) - \frac{h_{\mu\nu}}{2}(R^{(1)} + \frac{2}{\ell^2}h) \tag{4.10}$$

$$R^{(2)}_{\mu\nu} = \nabla_\lambda \Gamma^{(2)\lambda}_{\mu\nu} - \nabla_\nu \Gamma^{(2)\lambda}_{\mu\lambda} + \Gamma^{(1)\lambda}_{\lambda\sigma}\Gamma^{(1)\sigma}_{\mu\nu} - \Gamma^{(1)\lambda}_{\nu\sigma}\Gamma^{(1)\sigma}_{\mu\lambda} \tag{4.11}$$

$$\Gamma^{(2)\lambda}_{\mu\nu} = -\frac{h^{\lambda\sigma}}{2}(\nabla_\nu h_{\sigma\mu} + \nabla_\mu h_{\sigma\nu} - \nabla_\sigma h_{\mu\nu}). \tag{4.12}$$

The one-form $\mathcal{K}$ in (3.1) may now be constructed to second order from $E^{(1)}(h^{(2)})$ and shown to be the divergence of a two-form $\mathcal{F}(h^{(2)})$. The resulting boundary expression for the charges

$$Q_{boundary}(\xi) = -\frac{1}{16\pi G}\oint_{\partial\Sigma} *\mathcal{F}(h^{(2)}) \tag{4.13}$$

yields the expression quoted in (3.5). The bulk expression is then obtained by integrating by parts. When ξ is a background Killing vector it is straightforward to write this bulk charge explicitly

$$Q_{bulk}(\xi) = \frac{1}{16\pi G}\int_\Sigma *(\xi^\mu E^{(2)}_{\mu\nu}(h^{(1)})dx^\nu). \tag{4.14}$$

For general ξ one can write a similar but somewhat more complicated expression.

At this point we have not assumed Brown-Henneaux boundary conditions. We note that it is crucial that the $\partial_\rho h_{--}$ terms in (3.5) are included; these terms are omitted in some discussions in the literature. Without them the bulk and boundary expressions (4.14) and (4.13) would not be equal.

4.1 Chirality confinement

We now turn to a discussion of solutions of the linearized equations and their second-order back reaction. One may consider a basis of eigenmodes of $\xi^L_0 = \partial_-$ and $\xi^R_0 = \partial_+$, or equivalently energy and angular momentum. Such eigenmodes were constructed in [1], where it was shown that all the (non-gauge) modes obeying the boundary conditions (2.6) have vanishing left charges and are in the $(h_L, h_R) = (0, 2)$ highest

weight representation of $SL(2,\mathbb{R})_L \times SL(2,\mathbb{R})_R$. These are the right-moving boundary gravitons and can be constructed from non-trivial ξ^R_{-2} and ξ^R_{-1} diffeomorphisms on the AdS_3 background. There are also weight $(2,0)$ left-moving excitations, but these can be eliminated by trivial ξ^L_{-2} and ξ^L_{-1} diffeomorphisms. This is in contrast with the situation for generic μ where there are three types of eigenmodes in highest weight representations: chiral left and right boundary gravitons as well as massive gravitons transforming in a non-chiral highest weight $\frac{1}{2}(3+\mu\ell, -1+\mu\ell)$ representation. As $\mu\ell \to 1$, the weight of the massive graviton approaches $(2,0)$ and its wave function degenerates with that of the left-moving boundary graviton. Consequently it can also be eliminated by a trivial diffeomorphism. Hence the disappearance of the massive and left moving representations at the chiral point is a direct result of the enhancement of the group of trivial symmetries.

However, there is no guarantee that all solutions of the linearized equations obeying the boundary conditions (2.6) have an expansion in terms of (ξ^L_0, ξ^R_0) eigenmodes, or fall into highest weight representations. Interestingly, modes without such an expansion do exist. Examples were explicitly constructed in [8] (building on results of [10]) and will be denoted $h^{(1)}_{GKP}$. $h^{(1)}_{GKP}$ cannot be Fourier expanded as eigenmodes of ∂_τ because it grows linearly in τ. Moreover the GKP modes are non-chiral: the quadratic bulk expressions for left and right moving charges are non-zero

$$E_L \equiv Q^L_{bulk}(\xi^L_0, h^{(1)}_{GKP}) = -\frac{\ell}{12G}, \quad E_R \equiv Q^R_{bulk}(\xi^R_0, h^{(1)}_{GKP}) = -\frac{\ell}{24G}. \tag{4.15}$$

On the other hand, we can also compute the charges from the boundary expression. This involves first solving for the second order perturbation $h^{(2)}_{GKP}$ and then evaluating the boundary integral. Since the bulk and boundary expressions are equal we must have

$$E_L = Q^L_{boundary}(\xi^L_0, h^{(2)}_{GKP}) = -\frac{\ell}{12G}, \tag{4.16}$$

where we have imposed the condition (3.6). This cannot be nonzero if $h^{(2)}_{GKP}$ obeys the boundary condition (2.6). We conclude that $h^{(2)}_{GKP}$ violates the boundary condition, and there is no exact solution to chiral gravity already at second order with the prescribed boundary condition. Explicit expressions for $h^{(1)}_{GKP}$ and $h^{(2)}_{GKP}$ are given below. $h^{(2)}_{GKP}$ grows linearly at infinity so that $\partial_\rho h^{(2)}_{GKP}$ gives a nonvanishing contribution to the boundary expression for the left charge.

This resolves the apparent contradiction between the the vanishing of E_L and the existence of boundary-condition-obeying solutions of the linearized equations with

nonzero E_L. The latter are obstructed at second order and are not the linearization of boundary-condition-obeying solutions of the exact equations.

In the introduction we made an analogy between non-chiral solutions of linearized chiral gravity and free quark solutions of linearized QCD: neither are approximate finite-energy solutions of the exact theory. An alternate, purely classical, analogy can be found in Maxwell electromagnetism coupled to a charged scalar in 1+1 dimensions. At linear order there are scalar field configurations of order ϵ with finite charge and finite energy. However these disappear from the finite energy spectrum at quadratic order: there is an electric field of order ϵ^2 which carries infinite energy due to an infrared divergence. So there are no finite energy excitations with nonzero charge: charge is confined. Here we are finding in analogy that non-chiral excitations are confined. In the Maxwell case, there is only one conserved quantity – the electric charge – which must vanish. This implies that the linearized solutions must obey a one parameter constraint in order to approximate exact solutions to the theory. In chiral gravity there are infinitely many conserved charges Q^L which must vanish. This leads to infinitely many constraints, and the elimination of an entire (left) chiral sector of the theory.

We have shown that the linearization of all finite excitations of chiral gravity must be chiral in the sense that the quadratic bulk expression for E_L (as well as the other left charges) must vanish. This is irrelevant to the energy eigenmodes which are in any case chiral, but it eliminates the nonchiral mode $h^{(1)}_{GKP}$ which, from (4.15), has $E_L = -\frac{\ell}{12G}$. In principle there could be additional modes which are chiral but still have negative energy $E = E_L + E_R = E_R$. This would ruin perturbative stability. This seems highly unlikely since all linear chiral modes are associated with asymptotic symmetries, and we know already that the ASG is generated by exactly one copy of the Virasoro algebra. This Virasoro algebra is already accounted for by the $(0, 2)$ mode.

For the sake of completeness, in sections (4.2) and (4.3) we will compute explicitly the second order perturbation resulting from the various alleged counterexamples to the chiral gravity conjecture. We will see the infrared divergence described above and conclude that the linearization of the exact spectrum consists only of the right-moving boundary gravitons.

We note that it is in principle straightforward to find *all* solutions of the linearized constraint equations in global coordinates, rather than just the energy eigenmodes described above. However, the analogous computation has already been solved in Poincaré coordinates [7]. So we will work primarily in Poincaré coordinates. We will

then show in section (4.4) that on global AdS_3 all linearized solutions which are non-singular at second order must be chiral and obey the linearized Einstein equations.

4.2 The CDWW modes

Carlip, Deser, Waldron and Wise (CDWW) have described all solutions of the linearized equations of motion which are smooth in Poincaré coordinates [7]. These include many nonchiral modes. We will first show that all of these nonchiral modes are singular at second order on the boundary of the Poincaré patch. All modes which are nonsingular at second order are chiral and obey the linearized Einstein equation.

We use Poincaré coordinates[7]

$$ds^2 = \frac{-dt^+dt^- + dz^2}{z^2} \tag{4.17}$$

and light-front gauge

$$h^{(1)}_{--} = h^{(1)}_{+-} = h^{(1)}_{-z} = 0. \tag{4.18}$$

Following [7] we may integrate out $\partial_- h^{(1)}_{+z}$ and $\partial_-^2 h^{(1)}_{++}$ in the quadratic action. The equation of motion for $h^{(1)}_{zz}$ becomes

$$\partial_+\partial_- h^{(1)}_{zz} = \frac{1}{4z^2}[z^2\partial_z^2 + 3z\partial_z + (-\mu^2 + 4\mu - 3)]h^{(1)}_{zz} \tag{4.19}$$

The general solution of (4.19) is a real linear combination of the modes

$$h^{(1)}_{\omega^+,\omega^-} = \sqrt{\frac{\omega}{4\pi E}}\frac{1}{z}e^{-i(\omega^+t^+ + \omega^-t^-)}J_{|2-\mu|}(2\omega z) \tag{4.20}$$

$$h^{(1)*}_{\omega^+,\omega^-} = \sqrt{\frac{\omega}{4\pi E}}\frac{1}{z}e^{i(\omega^+t^+ + \omega^-t^-)}J_{|2-\mu|}(2\omega z) \tag{4.21}$$

where

$$\omega^2 = \omega^+\omega^-, \quad E = \frac{\omega^+ + \omega^-}{2}, \quad k = \frac{\omega^+ - \omega^-}{2}$$

A general solution can be written as a wave packet

$$h^{(1)}_{zz} = \int d\omega dk[a(\omega^+,\omega^-)h^{(1)}_{\omega^+,\omega^-} + a^*(\omega^+,\omega^-)h^{(1)*}_{\omega^+,\omega^-}] \tag{4.22}$$

[7] The orientation here is $\epsilon_{t\phi\rho} = \sqrt{-g}$, or equivalently, $\epsilon_{+-z} = \sqrt{-g}$.

The other components of the linear fluctuation are obtained from (4.22) by

$$\partial_- h^{(1)}_{+z} = \frac{1}{2}[\partial_z + \frac{-\mu+2}{z}]h^{(1)}_{zz} \tag{4.23}$$

$$\partial^2_- h^{(1)}_{++} = \frac{1}{2}[2\partial_+\partial_- - \frac{\mu}{z}\partial_z + \frac{\mu^2-3\mu}{z^2}]h^{(1)}_{zz} \tag{4.24}$$

The left moving charges Q^L can now be computed from the bulk quadratic expression (4.9). They are in general nonzero. For example

$$E_L = -\frac{1}{128\pi G}\int dzdx\left[z^3\left((\partial_z h^{(1)}_{zz})^2 + 4(\partial_- h^{(1)}_{zz})^2\right)\right. \tag{4.25}$$
$$\left. +\frac{1}{2}\,\partial_z\left(z^2(9+z\partial_z)(h^{(1)}_{zz})^2\right)\right]$$
$$= -\frac{1}{128\pi G}\int dzdx\left[4z^3\left(-\partial_+\partial_- h^{(1)}_{zz}h^{(1)}_{zz} + (\partial_- h^{(1)}_{zz})^2\right]\right) \tag{4.26}$$
$$\left. +\frac{1}{2}\partial_z\left(z^2(9+z\partial_z)(h^{(1)}_{zz})^2\right)\right]$$

We have discarded here total derivatives of t^- which vanish upon integration over x. This expression is a total derivative plus a negative semi-definite term.

$$E_L = -\frac{1}{32\pi G}\int dzdxz^3\left[-\partial_+\partial_- h^{(1)}_{zz}h^{(1)}_{zz} + (\partial_- h^{(1)}_{zz})^2\right] \tag{4.27}$$

$$= -\frac{1}{64\pi G}\int d\omega dk\ \ \omega^-\ |a(\omega^+,\omega^-)|^2 \tag{4.28}$$

This vanishes if and only the mode has support in the region

$$w^- = 0, \tag{4.29}$$

In this case $h^{(1)}$ obeys the linearized Einstein equation

$$\mathcal{G}^{(1)}_{\mu\nu}(h^{(1)}) = 0. \tag{4.30}$$

In order to make this completely explicit, we will now compute the second order perturbation of the CDWW modes. We will compute the curvature rather than the metric, as divergences in the latter can be coordinate artifacts. The $--$ component of (4.4) is

$$\partial_z\mathcal{G}^{(1)}_{--}(h^{(2)}) - \partial_-\mathcal{G}^{(1)}_{z-}(h^{(2)}) \tag{4.31}$$

$$= \frac{1}{2}z^3\left[\partial_- h^{(1)}_{zz}\partial_- h^{(1)}_{zz} + \partial_-\left(2z\partial_-(h^{(1)}_{+z}\partial_- h^{(1)}_{zz}) - 2z h^{(1)}_{zz}\partial_-^2 h^{(1)}_{+z} - \partial_- h^{(1)}_{zz}(z\partial_z + \frac{5}{2})h^{(1)}_{zz}\right)\right].$$

A boundary condition for this differential equation is obtained by noting that at the boundary point $z = \infty$, (2.6) implies $\mathcal{G}^{(1)}_{--}(h^{(2)}; \infty, t^+, t^-) = 0$. If $h^{(1)}$ is one of the ω_+, ω_- eigenmodes, equation (4.31) decomposes into three equations which depend on $t^\pm$ as either $e^{\pm 2i(\omega_+ t^+ + \omega_- t^-)}$ or are constant in $t^\pm$. Consider the constant piece, for which $\mathcal{G}^{(1)}_{z-}(h^{(2)})$ vanishes by symmetry. We may then solve for the constant part of $\mathcal{G}^{(1)}_{--}(h^{(2)})$:

$$\mathcal{G}^{(1)}_{--}(h^{(2)}; z, t^+, t^-) = -\frac{1}{2}\omega_-^2 \int_{-\infty}^{z} dz' z'^3 h^{(1)*}_{\omega^+,\omega^-} h^{(1)}_{\omega^+,\omega^-} \tag{4.32}$$

which is strictly negative unless $\omega^- = 0$. Thus it is impossible for $\mathcal{G}^{(1)}_{--}(h^{(2)})$ to vanish everywhere on the boundary $z = 0$ and $z = \infty$ unless $\partial_- h^{(1)}$ itself vanishes everywhere. This leaves only the chiral $\omega^- = 0$ modes which solve the linearized Einstein equation. We see explicitly that the linearized modes or, since the right hand side is always negative, any superposition thereof must obey the linearized Einstein equation. Looking at the Fourier modes of $\mathcal{G}^{(1)}_{--}(h^{(2)})$ gives more constraints leading again to $\omega^- = 0$.

We note that the above expressions for the curvature at second order, and hence the conclusion that the boundary conditions are violated, follows directly from the perturbative expansion of the equations of motion. Thus although our discussion was motivated by charge conservation, our conclusions ultimately do not rely on any particular expressions for or properties of the charges.

4.3 The GKP mode

A interesting nonchiral solution of the linearized equations was constructed by Giribet, Kleban and Porrati (GKP) in [9]. This mode is not an (ξ_0^L, ξ_0^R) eigenmode but nevertheless obeys the Brown-Henneaux boundary conditions (2.6). It may be written as

$$h^{(1)GKP}_{\mu\nu} = \mathcal{L}^R_{-1}(y(\tau,\rho)\mathcal{L}^L_{-2}\bar{g}_{\mu\nu}) + \mathcal{L}_\xi \bar{g}_{\mu\nu} \tag{4.33}$$

where $\mathcal{L}^{L,R}_n$ is the Lie derivative with respect to $\xi^{L,R}_n$ and

$$y(\tau,\rho) = -i\tau - \ln(\cosh\rho) \tag{4.34}$$

$$\xi = -\frac{iy(\tau,\rho)\sinh(\rho)}{2\ell^2\cosh^5(\rho)} e^{-i(\tau^+ + 2\tau^-)}\xi_0^R. \tag{4.35}$$

The conserved charges are

$$E^L_{GKP} = -\frac{\ell}{12G} \tag{4.36}$$

$$E^R_{GKP} = -\frac{\ell}{24G} \tag{4.37}$$

We may now solve asymptotically for $h^{(2)GKP}$ using (4.4), which reduces to

$$E^L = \frac{1}{16\pi\ell G}\oint d\phi(2\partial_\rho h^{(2)}_{--} - \partial^2_\rho h^{(2)}_{--}) = E^L_{GKP} \tag{4.38}$$

$$E^R = \frac{1}{16\pi\ell G}\oint d\phi(4h^{(2)}_{++} - 4\partial_\rho h^{(2)}_{++} + \partial^2_\rho h^{(2)}_{++}) = E^R_{GKP} \tag{4.39}$$

The ϕ independent solution is

$$h^{(2)}_{--} = 4G\ell E^L\rho + \cdots \tag{4.40}$$

$$h^{(2)}_{++} = 2G\ell E^R + \cdots \tag{4.41}$$

where $\ldots$ denotes terms which are subleading in ρ. In particular, for the GKP modes, we have

$$h^{(2)GKP}_{--} = -\frac{\ell^2\rho}{3} + \cdots \tag{4.42}$$

$$h^{(2)GKP}_{++} = -\frac{\ell^2}{12} + \cdots \tag{4.43}$$

From (4.42) we see that the Brown-Henneaux boundary conditions (2.6) are violated. We conclude that $h^{(1)GKP}$ is not the linearization of an exact solution to the equations of motion.

4.4 Global modes

We can now argue that *all* solutions to the linearized equations of motion that obey Brown-Henneaux boundary conditions at second order must be solutions of the linearized Einstein equations. In particular, we are left only with the right-moving boundary gravitons.

To prove this, one could study the linearized equations of motion in global rather than Poincaré coordinates. However this can be avoided by noting that every mode which is smooth and asymptotically AdS_3 in global coordinates is smooth on the

Poincare patch and hence has an expansion in CDWW modes of section (4.2).[8] We have seen that of these modes only the ones with vanishing Einstein tensor obey Brown-Henneaux boundary conditions at second order. As the right hand side of (4.32) is negative definite, we cannot cancel this divergence for any linear superpositions of modes. Hence all global modes must obey the linearized Einstein equation.

5 A Birkhoff-like theorem

We have seen that any solution of chiral gravity is, at the linearized level, locally AdS_3. This might lead one to suspect that all solutions of chiral gravity are locally AdS_3 at the full non-linear level. In this section we will see that this is indeed the case for a particularly simple class of solutions: those which are stationary and axially symmetric. For this class of solutions the full non-linear equations of motion, although still surprisingly complicated, are reasonably tractable. We will conclude that, once we impose Brown-Henneaux boundary conditions, the only solutions are the BTZ black holes.

A similar result was obtained for general values of μ by [46, 47], who made the somewhat stronger assumption of a hypersurface orthogonal Killing vector field.

5.1 Stationarity and axial symmetry

We start by studying the equations of motion of TMG for stationary, axially symmetric spacetimes, following the approach of [48].

A three dimensional spacetime with two commuting $U(1)$ isometries may, through judicious choice of coordinates, be written in the form

$$ds^2 = -X^+(\sigma)d\tau^2 + X^-(\sigma)d\phi^2 + 2Y(\sigma)dtd\phi + \frac{d\sigma^2}{X^+(\sigma)X^-(\sigma) + Y(\sigma)^2} \tag{5.1}$$

The two $U(1)$ isometries are the generated by Killing vectors ∂_τ and ∂_ϕ. We are interested in axially symmetric solutions, so we will take the angular direction to be periodic $\phi \sim \phi + 2\pi$. We have chosen the coefficient of $d\sigma^2$ for future convenience.

The geometry of the solution is encoded in the three functions $X^\pm(\sigma), Y(\sigma)$, which we will package into a three dimensional vector $\mathbf{X}$ with components X^i, $i = 0, 1, 2$

[8] Of course, the converse is not true: modes which are well behaved on the Poincare patch may not be well-behaved globally.

given by

$$X^0 = \frac{X^+(\sigma) - X^-(\sigma)}{2}, \qquad X^1 = \frac{X^+(\sigma) + X^-(\sigma)}{2}, \qquad X^2 = Y \tag{5.2}$$

The dynamics of stationary, axial metrics in three dimensions may be thought of as the dynamics of a particle with position $\mathbf{X}(\sigma)$ moving in the auxiliary space $\mathbb{R}^{2,1}$ parameterized by $\mathbf{X}$.

For the ansatz (5.1), the equations of motion of TMG are

$$-2\mu\mathbf{X}'' = 2\mathbf{X}\times\mathbf{X}''' + 3\mathbf{X}'\times\mathbf{X}'' \tag{5.3}$$

$$4 = \mathbf{X}'^2 - \frac{2}{\mu}\mathbf{X}\cdot(\mathbf{X}'\times\mathbf{X}'') \tag{5.4}$$

We have set $\ell = 1$ for convenience. Here $'$ denotes ∂_σ and we have defined the Lorentz invariant dot product and cross product[9]

$$\mathbf{A}\cdot\mathbf{B} \equiv \eta_{ij}A^iB^j, \qquad \text{and} \qquad (\mathbf{A}\times\mathbf{B})^i \equiv \eta^{il}\epsilon_{ljk}A^jB^k, \quad \epsilon_{012} = 1 \tag{5.5}$$

In order to understand these equations, it is helpful to note that for our ansatz (5.1) the action of TMG is

$$I = \frac{1}{16\pi G}\int d\sigma \frac{1}{2}\left(\mathbf{X}'^2 - \frac{1}{\mu}\mathbf{X}\cdot(\mathbf{X}'\times\mathbf{X}'')\right) \tag{5.6}$$

This is the action of a Lorentz invariant particle mechanics in $\mathbb{R}^{2,1}$. Equation (5.3) is found by varying this reduced action with respect to $\mathbf{X}$. Equation (5.4) represents an additional constraint which arises due to gauge-fixing. In fact, the right hand side of (5.4) is just the conserved Hamiltonian of the reduced action (5.6).

The equation of motion (5.3) can easily be integrated once. To see this, note that tha action (5.6) is invariant under Lorentz transformations in $\mathbb{R}^{2,1}$. Hence there is a conserved angular momentum $\mathbf{J}$ associated to these Lorentz transformations, which we can compute using the Noether procedure (taking into account the higher derivative terms)

$$\mathbf{J} = \frac{1}{16\pi G}\left(\mathbf{X}\times\mathbf{X}' - \frac{1}{2\mu}[\mathbf{X}'\times(\mathbf{X}\times\mathbf{X}') - 2\mathbf{X}\times(\mathbf{X}\times\mathbf{X}'')]\right) \tag{5.7}$$

[9] In Lorentzian signature some of the usual cross product identities must be altered (e.g. $\mathbf{A}\times(\mathbf{B}\times\mathbf{C}) = \mathbf{C}(\mathbf{A}\cdot\mathbf{B}) - \mathbf{B}(\mathbf{A}\cdot\mathbf{C})$ differs by a sign from the usual case).

One can check explicitly that (5.7) is the the first integral of (5.3). The dynamics of the system is given by the set of second order differential equations (5.7) and (5.4). The auxiliary angular momentum $\mathbf{J}$ should not be confused with the physical angular momentum of the spacetime, although we shall see that they are closely related. With the help of (5.3) and (5.4) we can write (5.7) as a second order differential equation for $\mathbf{X}''$:

$$2\mathbf{X}^2\mathbf{X}'' = 32\pi G\mu\mathbf{J} - 2\mu\mathbf{X}\times\mathbf{X}' + \mathbf{X}'(\mathbf{X}\cdot\mathbf{X}') + \mathbf{X}(6 - \frac{5}{2}\mathbf{X}'^2) \tag{5.8}$$

This equation can then be integrated to give a solution $\mathbf{X}(\sigma)$ to the equations of motion. For future reference, we note that with the help of the equations of motion we can write the Hamiltonian constraint as

$$(\mathbf{X}'^2 - 4) = \frac{2}{\mu}\mathbf{X}\cdot(\mathbf{X}'\times\mathbf{X}'') = -\frac{4}{3}\mathbf{X}\cdot\mathbf{X}'' \tag{5.9}$$

5.2 Boundary conditions

We will now consider axially symmetric, stationary solutions of TMG obeying Brown-Henneaux boundary conditions.

We start by noting that global metric on AdS_3 can be written as

$$ds^2 = -(2\sigma+1)d\tau^2 + 2\sigma d\phi^2 + \frac{d\sigma^2}{2\sigma(2\sigma+1)} \tag{5.10}$$

which is of the form (5.1) with

$$\mathbf{X}_{AdS} = 2\sigma(0,1,0) + \frac{1}{2}(1,1,0) \tag{5.11}$$

Here σ is related to the usual global radial coordinate on AdS_3 by $\sigma = \frac{1}{2}\sinh^2\rho$. The asymptotic boundary is at $\sigma\to\infty$. Likewise, the BTZ black hole can be written in the form (5.1) with

$$\mathbf{X}_{BTZ} = 2\sigma(0,1,0) - 4GM(1,1,0) + 4GJ(0,0,1) \tag{5.12}$$

Here M and J are the ADM mass and angular momentum of the BTZ black hole in Einstein gravity. As we are working in units with $\ell = 1$, empty AdS_3 has energy $M = -1/8G$.

Let us now consider an arbitrary metric obeying Brown-Henneaux boundary conditions. Comparing with the AdS metric (5.10) one can check that a metric of the

form (5.1) obeys Brown-Henneaux boundary conditions if

$$\mathbf{X} = 2\sigma(0,1,0) + \mathcal{O}(1), \qquad \text{as} \quad \sigma \to \infty \tag{5.13}$$

In analogy with (5.12) we will write this boundary condition as

$$\mathbf{X} = 2(\sigma - \sigma_0)(0,1,0) - 4GM(1,1,0) + 4GJ(0,0,1) + \dots \tag{5.14}$$

where σ_0, M and J are constants, G is Newton's constant and $\dots$ denotes terms which vanish as $\sigma \to \infty$. By comparing with (5.12), we see that M and J are the usual ADM mass and angular momentum of the spacetime as measured at asymptotic infinity in Einstein gravity. The parameter σ_0 is just a shift in the radial coordinate and does not have a coordinate-independent meaning.

Let us now consider solutions to the equations of motion of TMG with the boundary conditions (5.14). As the angular momentum $\mathbf{J}$ is a constant of motion we can compute it at $\sigma \to \infty$. Plugging (5.14) into (5.7) we find

$$2\pi\ \mathbf{J} = (J, 0, -M) + \frac{1}{\mu}(-M, 0, J) \tag{5.15}$$

The auxiliary angular momentum $\mathbf{J}$ is just a rewriting of the usual mass and angular momentum of the solution. We note that

$$(2\pi\mathbf{J})^2 = \left(1 - \frac{1}{\mu^2}\right)\left(M^2 - J^2\right) \tag{5.16}$$

For values of $\mu > 1$, we see that the angular momentum vector is spacelike for solutions obeying the cosmic censorship bound $M > J$. For extremally rotating solutions $\mathbf{J}$ is null, $\mathbf{J}^2 = 0$.

Finally, we turn to the case of chiral gravity ($\mu = 1$) with Brown-Henneaux boundary conditions. In this case the angular momentum vector is always null:

$$2\pi\mathbf{J} = (J - M)(1, 0, 1), \qquad \mathbf{J}^2 = 0 \tag{5.17}$$

This property will turn out to be very useful.

5.3 Solutions

We will now specialize to chiral gravity ($\mu = 1$) and study axially symmetric, stationary solutions obeying Brown-Henneaux boundary conditions. We will demonstrate

that if the spacetime has a single asymptotic boundary obeying Brown-Henneaux boundary conditions, then the spacetime must be locally AdS_3. We will also assume that $\mathbf{X}(\rho)$ is an analytic function of ρ.

At $\sigma \to \infty$, the vector $\mathbf{X}$ is spacelike. As long as $\mathbf{X}$ remains spacelike, we can continue to smoothly evolve our metric into the interior using (5.8). In fact, $\mathbf{X}$ must become null – with $\mathbf{X}^2 = 0$ – for some finite value of σ. To see this, consider what would happen if $\mathbf{X}^2$ remained strictly positive for all values of σ. In this case the evolution equation (5.8) would allow us to evolve $\mathbf{X}$ all the way to $\sigma \to -\infty$. The region $\sigma \to -\infty$ then represents an additional asymptotic boundary. To prove this, note that in order for $\mathbf{J}$ to remain finite at $\sigma \to -\infty$, $\mathbf{X}^2$ must either remain finite or diverge no more quickly than σ^2. The line element

$$ds^2 \sim \frac{d\sigma^2}{\mathbf{X}^2} + \dots \tag{5.18}$$

then implies that points with $\sigma \to -\infty$ lie an infinite proper distance from points with finite σ. As we are assuming Brown-Henneaux boundary conditions with a *single* asymptotic boundary, we must not allow this additional boundary at $\sigma \to -\infty$. We conclude that there must be finite value of σ where $\mathbf{X}$ becomes null, i.e. $\mathbf{X}^2 = 0$.

We will now proceed to study the equation of motion near the point where $\mathbf{X}^2 = 0$. First, let us shift our coordinate σ so that this point occurs at $\sigma = 0$. We will assume that the metric is analytic at this point, so is equal to its Taylor expansion

$$\mathbf{X} = \sum_{n \geq 0} \frac{1}{n!} \sigma^n \mathbf{X}_n \tag{5.19}$$

The coefficients $\mathbf{X}_n$ are finite and given by derivatives of $\mathbf{X}$ at $\sigma = 0$, with $\mathbf{X}_0^2 = 0$. In fact, one can check that the point $\sigma = 0$ is either a horizon or an origin of polar coordinates, depending on the relative values of $\mathbf{X}_0$ and $\mathbf{X}_1$.

We now turn to the equations of motion. By plugging (5.19) into (5.8) and expanding order by order in powers of σ we obtain a set of recursion relations which determine the $\mathbf{X}_n$ in terms of $\mathbf{X}_0$ and $\mathbf{X}_1$. We will now proceed to show that these recursion relations imply that all the terms with $n \geq 2$ in the Taylor expansion (5.19) vanish. This will imply that our Taylor expansion converges, hence the solution can be smoothly matched on to the metric at infinity. Indeed, comparing with equation (5.12) we see that our solution

$$\mathbf{X} = \mathbf{X}_0 + \sigma \mathbf{X}_1 \tag{5.20}$$

is simply the BTZ black hole.[10] This allows us to conclude that our solution is locally AdS_3.

In order to demonstrate this, let us now expand equation (5.8) order by order in σ. The order σ^0 term just fixes the angular momentum vector

$$32\pi G\mathbf{J} = 2\mathbf{X}_0 \times \mathbf{X}_1 - \mathbf{X}_1(\mathbf{X}_0 \cdot \mathbf{X}_1) - \mathbf{X}_0(6 - \frac{5}{2}\mathbf{X}_1^2) \tag{5.21}$$

In chiral gravity, $\mathbf{J}$ must be null, so that

$$(32\pi G\mathbf{J})^2 = -4(\mathbf{X}_0 \cdot \mathbf{X}_1)^2(\mathbf{X}_1^2 - 4) = 0 \tag{5.22}$$

implying that either $\mathbf{X}_1^2 - 4$ or $\mathbf{X}_0 \cdot \mathbf{X}_1$ vanish. We will consider the following cases separately:

Case 0: $\boldsymbol{X}_0 = 0$

In this case $\mathbf{J} = 0$ and it is easy to prove directly that all the higher order terms vanish. In particular, (5.7) implies that

$$\mathbf{J} \cdot \mathbf{X} = \mathbf{X}^2\mathbf{X}'^2 - (\mathbf{X} \cdot \mathbf{X}')^2 = 0 \tag{5.23}$$

so that $\mathbf{X} \times \mathbf{X}'$ is null. We also see that

$$\mathbf{J} \cdot (\mathbf{X} \times \mathbf{X}' - 2\mathbf{X}) = \mathbf{X}^2(\mathbf{X}'^2 - 4) = 0 \tag{5.24}$$

so that $\mathbf{X}'^2 = 4$. Thus in the region where $\mathbf{X}^2$ is positive, $\mathbf{X}$ and $\mathbf{X}'$ are spacelike vectors whose cross product $\mathbf{X} \times \mathbf{X}'$ is null. One can use this condition to show that $\mathbf{X}$ and $\mathbf{X}'$ obey[11]

$$\mathbf{X} \times \mathbf{X}' = -\sqrt{\mathbf{X}'^2}\mathbf{X} + \sqrt{\mathbf{X}^2}\mathbf{X}' \tag{5.25}$$

Plugging these identities into (5.8) we conclude that $\mathbf{X}'' = 0$. Thus all of the higher order terms in the Taylor expansion vanish and the solution is just the BTZ black hole.

[10] In comparing (5.20) with (5.12) we must remember that in (5.20) we have shifted σ to put the horizon at $\sigma = 0$, in contrast to (5.12).

[11] For any two spacelike vectors $\mathbf{A}$ and $\mathbf{B}$ whose cross product is null one has the identity $\mathbf{A} \times \mathbf{B} = \pm\sqrt{\mathbf{A}^2}\mathbf{B} \pm \sqrt{\mathbf{B}^2}\mathbf{A}$ where the signs depend on the relative orientations of the vectors. We have fixed the signs here by comparing to the behavior at asymptotic infinity (where $\mathbf{X}$ approaches that of an extremally rotating BTZ black hole with $M = J$).

Case 1: $\boldsymbol{X}_1^2 = 4$ *and* $\boldsymbol{X}_0 \cdot \boldsymbol{X}_1 \neq 0$

In this case we must work a little harder and examine the terms in the Taylor expansion (5.19) term by term. The equation for $\mathbf{X}_2$ is found by expanding (5.8) to linear order in σ:

$$3\mathbf{X}_0 \cdot \mathbf{X}_1\mathbf{X}_2 + 2\mathbf{X}_0 \times \mathbf{X}_2 - \mathbf{X}_1\mathbf{X}_0 \cdot \mathbf{X}_2 + 5\mathbf{X}_0\mathbf{X}_1 \cdot \mathbf{X}_2 = 0 \tag{5.26}$$

The Hamiltonian equation expanded to order σ^0 is

$$\mathbf{X}_0 \cdot \mathbf{X}_2 = 0 \tag{5.27}$$

In fact, $\mathbf{X}_2 = 0$ is the only solution to this equation. To see this, note that since $\mathbf{X}_0 \cdot \mathbf{X}_1 = 0$ the vectors $\mathbf{X}_0$, $\mathbf{X}_1$ and $\mathbf{X}_0 \times \mathbf{X}_1$ form a basis for $\mathbb{R}^{2,1}$. So we may expand

$$\mathbf{X}_2 = a\mathbf{X}_0 + b\mathbf{X}_1 + c\mathbf{X}_0 \times \mathbf{X}_1 \tag{5.28}$$

for some constants a, b, c. Plugging into the equations of motion we find that each of these constants must vanish, so $\mathbf{X}_2 = 0$.

We will now prove by induction that all of the higher order terms in the expansion (5.19) must vanish as well. Let us start by assuming that all of the quadratic and higher terms in the expansion (5.19) vanish up to a given order m. That is, let us assume that

$$\mathbf{X} = \mathbf{X}_0 + \sigma\mathbf{X}_1 + \sum_{n \geq m} \frac{1}{n!}\sigma^n \mathbf{X}_n \tag{5.29}$$

for some $m \geq 2$. Expanding equation (5.8) to order σ^{m-1} gives

$$(4m-5)\mathbf{X}_0 \cdot \mathbf{X}_1\mathbf{X}_m + 2\mathbf{X}_0 \times \mathbf{X}_m - \mathbf{X}_0 \cdot \mathbf{X}_m\mathbf{X}_1 + 5\mathbf{X}_1 \cdot \mathbf{X}_m\mathbf{X}_0 = 0 \tag{5.30}$$

and the Hamiltonian constraint at order σ^{m-2} gives

$$\mathbf{X}_0 \cdot \mathbf{X}_m = 0 \tag{5.31}$$

Expanding $\mathbf{X}_m$ in the basis as above, one can again show that each term in the basis expansion vanishes separately. Hence $\mathbf{X}_m$ vanishes and the inductive hypothesis (5.29) holds up to order $m+1$. In the previous paragraph we proved the case $m = 2$, so by induction it follows that that *all* $\mathbf{X}_n$, $n \geq 2$ must vanish.

Case 2: $\boldsymbol{X}_0 \neq 0$ *but* $\boldsymbol{X}_0 \cdot \boldsymbol{X}_1 = 0$

This special case is a bit more complicated. We note that since $\mathbf{X}_0$ is null and $\mathbf{X}_0 \cdot \mathbf{X}_1 = 0$ it follows that

$$\mathbf{X}_0 \times \mathbf{X}_1 = \pm\sqrt{\mathbf{X}_1^2}\mathbf{X}_0 \tag{5.32}$$

Expanding the Hamiltonian constraint to order σ^0 gives

$$\mathbf{X}_1^2 - 4 = 2\mathbf{X}_2 \cdot (\mathbf{X}_0 \times \mathbf{X}_1) = \pm 2\sqrt{\mathbf{X}_1^2}\mathbf{X}_2 \cdot \mathbf{X}_0 \tag{5.33}$$

Comparing to the second form of the Hamiltonian constraint

$$\mathbf{X}_1^2 - 4 = -\frac{4}{3}\mathbf{X}_2 \cdot \mathbf{X}_0 \tag{5.34}$$

we see that either $\mathbf{X}_1^2 = 4/9$ or $\mathbf{X}_1^2 = 4$. We will consider these two cases separately.

If $\mathbf{X}_1^2 = 4$ then we can show that all higher order terms in the Taylor expansion vanish. The Hamiltonian constraint is $\mathbf{X}_2 \cdot \mathbf{X}_0 = 0$. Expanding the equation of motion to linear order in σ gives equation (5.26), which in this case has the solution

$$\mathbf{X}_2 = a\mathbf{X}_0 \tag{5.35}$$

for some constant a. However, we find that the equation at order σ^3 implies that $a = 0$ so that $\mathbf{X}_2 = 0$. Likewise, equation (5.30) as a solution of the form

$$\mathbf{X}_m = a_m\mathbf{X}_0 \tag{5.36}$$

but the coefficient a_m is set to zero by the equation of motion at order σ^{m+1}. Proceeding in this manner we conclude that all of the $\mathbf{X}_n$ must vanish for $n \geq 2$, so the solution is just the BTZ black hole.

Finally, let us consider the case where $\mathbf{X}_1^2 = 4/9$. In this case the Hamiltonian constraint

$$\mathbf{X}_2 \cdot \mathbf{X}_0 = \frac{8}{3} \tag{5.37}$$

implies that $\mathbf{X}_2$ is non-zero. Aside from this small difference, the argument proceeds exactly as above. Expanding the equations of motion order by order in σ we discover that all of the terms in the Taylor expansion vanish except for $\mathbf{X}_0$, $\mathbf{X}_1$ and $\mathbf{X}_2$, which

obey

$$\mathbf{X}_0^2 = \mathbf{X}_2^2 = \mathbf{X}_0 \cdot \mathbf{X}_1 = \mathbf{X}_1 \cdot \mathbf{X}_2 = 0, \quad \mathbf{X}_1^2 = 4, \quad \mathbf{X}_0 \cdot \mathbf{X}_2 = \frac{8}{3}, \quad \mathbf{X}_0 \times \mathbf{X}_2 = \pm 4 \mathbf{X}_1. \tag{5.38}$$

One may check that the resulting solution

$$\mathbf{X} = \mathbf{X}_0 + \mathbf{X}_1 \rho + \frac{1}{2} \mathbf{X}_2 \rho^2 \tag{5.39}$$

is warped AdS_3. This solution does not obey Brown-Henneaux boundary conditions, because of the $\mathcal{O}(\rho^2)$ behavior at asymptotic infinity.

This completes the proof.

6 Log gravity

In this section we consider log gravity, which differs from chiral gravity in that the boundary conditions are relaxed to allow certain types of growth linear in ρ (and logarithmic in the proper radius) at infinity. The solutions of log gravity will have energies which are unbounded below as well as unbounded above. Nevertheless, the theory is of considerable interest as it contains a novel and mathematically natural class of solutions [34, 13, 21, 35] excluded from chiral gravity. Here we will show that log gravity is consistent insofar as the expressions for the conserved charges are finite. However, the left charges are in general nonzero, so log gravity is not chiral. This result is in agreement with [29].[12] Moreover, we shall see that log gravity contains within it a decoupled superselection sector which is identical to chiral gravity. The relation between this chiral $Q^L = 0$ subsector and the full spectrum of log gravity is reminiscent of the relation between the physical states of a first-quantized string (or any 2d gravity theory) and the larger Hilbert space including longitudinal modes and ghosts. Indeed, logarithmic CFTs appeared in the 2D gravity context in [49].

6.1 Boundary conditions and non-chirality

The starting point for the development of log gravity was the observation by Grumiller and Johansson (GJ) in that a solution to TMG at the chiral point can be obtained as [10]

$$h^{GJ} \equiv \lim_{\mu\ell \to 1} \frac{h^\mu - h^L}{\mu\ell - 1}. \tag{6.1}$$

[12] However, this result is not in agreement with reference [33], which neglected a term in the generators.

Here h^μ and h^L are the wave functions for the massive and left-moving gravitons, respectively. The mode h^{GJ} is a solution of the linearized equations of motion, but it is not an energy eigenstate and was not considered in [1]. Moreover h^{GJ} does not obey the Brown-Henneaux boundary conditions as certain components grow linearly in ρ at the boundary.

GJ then proposed that the Brown-Henneaux boundary conditions (2.6) be relaxed to allow metric fluctuations to grow linearly as $\rho \to \infty$ [10]. The mode h^{GJ} would be included in the spectrum of such a theory. However, this proposal does not lead to a consistent theory, because for the general such asymptotic perturbation the right-moving charge Q^R is linearly divergent and hence ill-defined. A modified approach [15] is to impose chiral boundary conditions for which h_{--} is allowed to grow linearly in ρ but h_{++} or h_{+-} are not. Specifically, we take [29, 33]

$$\begin{pmatrix} h_{++} = \mathcal{O}(1) & h_{+-} = \mathcal{O}(1) & h_{+\rho} = \mathcal{O}(e^{-2\rho}) \\ & h_{--} = \mathcal{O}(\rho) & h_{-\rho} = \mathcal{O}(\rho e^{-2\rho}) \\ & & h_{\rho\rho} = \mathcal{O}(e^{-2\rho}) \end{pmatrix} \tag{6.2}$$

The allowed diffeomorphisms are

$$\xi^+ = \epsilon^+(x^+) + 2e^{-2\rho}\partial_-^2\epsilon^-(x^-) + \cdots \tag{6.3}$$

$$\xi^- = \epsilon^-(x^-) + 2e^{-2\rho}\partial_+^2\epsilon^+(x^+) + \cdots \tag{6.4}$$

$$\xi^\rho = -\frac{1}{2}\left(\partial_+\epsilon^+(x^+) + \partial_-\epsilon^-(x^-)\right) + \cdots \tag{6.5}$$

The leading terms give two sets of Virasoro generators. The subleading terms are trivial and do not appear in the charges. It is straightforward to see that both Q^L and Q^R are finite for all elements of the ASG. However, since $\partial_\rho h_{--} \neq 0$ we find that

$$Q^L \neq 0, \tag{6.6}$$

so that log gravity is *not* chiral.

This opens up the possibility [10] that log gravity could be holographically dual to a logarithmic CFT. In fact, a logarithmic CFT can never be chiral.[13] Moreover it is not hard to see that the mode (6.1) lies in an indecomposable Virasoro representation (containing the left-moving highest weight representation of massless gravitons) characteristic of a logarithmic CFT [50].

[13] We thank M. Gaberdiel for pointing this out.

While we have seen that the charges are finite for log gravity, more work must be done to show that they actually generate the asymptotic symmetry group, or indeed if log gravity has a canonical formulation at all. A canonical formulation requires the construction of a closed invertible symplectic form Ω, or equivalently a Dirac bracket, on the physical phase space, The Dirac bracket is nonlocal in space and its construction involves inverting the constraints. Whether or not the constraints can be inverted depends on the boundary conditions, and so there is in general no guarantee that Dirac brackets exist for any boundary condition. Since there are physical zero norm states in log gravity, invertibility is not manifest. An elegant covariant construction of Ω was given for general relativity in [51] in the form of an integral $\Omega = \int d\Sigma_\alpha J^\alpha$ over a spatial slice, with $J^\alpha = \delta\Gamma^\alpha_{\nu\lambda} \wedge [\delta g^{\nu\lambda} + \frac{1}{2} g^{\nu\lambda}\delta \ln g] - \delta\Gamma^\lambda_{\nu\lambda}[\delta g^{\alpha\nu} + \frac{1}{2} g^{\alpha\nu}\delta \ln g]$ and δ the exterior derivative on the phase space. For TMG, there is an additional term proportional to $\frac{1}{\mu}\epsilon^{\alpha\lambda\nu}\delta\Gamma^\sigma_{\lambda\rho} \wedge \delta\Gamma^\rho_{\nu\sigma}$. It would be interesting to see by direct computation if this symplectic form is both finite and invertible for log gravity.

6.2 Decoupling the chiral gravity superselection sector

Log gravity in and of itself does not seem of so much interest because it is not unitary. Nonunitary theories of quantum gravity are generally easy to construct, and are not expected to shed much led on the presumably unitary theory which describes our four dimensional world. What makes log gravity interesting is that it contains chiral gravity, which has the possibility of being unitary, within it. The structure of this embedding is intriguing and could be useful for a full understanding of chiral gravity. In this section we explain how this embedding works.

Let Q^L_n denote the left Virasoro charges. The classical computation of the central charge is insensitive to the boundary conditions as long as the charges are well defined. Therefore the Dirac bracket algebra

$$\{Q^L_m, Q^L_n\} = i(m-n)Q^L_{m+n}, \tag{6.7}$$

has

$$c_L = 0 \tag{6.8}$$

as is the case for chiral gravity. The charges Q^L_n are conserved for all n. Therefore we can consistently truncate to the charge superselection sector of the theory with

$Q_n^L = 0$. Q_n^L are the Fourier transforms of the linearly growing terms in h

$$\partial_\rho h_{--} = 4\ell G \sum_n Q_n^L e^{in\phi}. \tag{6.9}$$

Therefore in the $Q_n^L = 0$ superselection sector we have

$$\partial_\rho h_{--} = 0. \tag{6.10}$$

This condition reduces the log gravity boundary conditions (6.2) to the chiral gravity boundary conditions (2.6).[14] Therefore the $Q_n^L = 0$ superselection sector of log gravity is precisely chiral gravity. Charge conservation guarantees that time evolution preserves the chiral boundary conditions and chiral gravity completely decouples from log gravity. Note that this result is nonperturbative.

At the classical level, this shows that solutions of chiral gravity cannot smoothly evolve into geometries with logarithmic behavior at infinity. Of course, we have not proven cosmic censorship so we cannot rule out singularities on the boundary for either log or chiral gravity.

One may phrase the issue of classical decoupling of chiral gravity in a different way in perturbation theory, where one can see the decoupling by direct computation without invoking charge conservation. If we excite two linearized modes of chiral gravity, will a log mode be excited at the next order? Do the chiral modes source the log modes? This question has already been answered to second order in our perturbative analysis of section 4. It is immediate from inspection of (4.31)-(4.32) that if we take $h^{(1)}$ to solve the linearized Einstein equation, then at second order $\mathcal{G}^{(2)}$ vanishes. Of course one can always add a log mode obeying the homogenous equation at second order, but as this is not required the log mode can be decoupled.

This analysis can be extended to all orders. If $h^{(1)}$ is nontrivial and is a linearized solution of chiral gravity, it also solves the linearized Einstein equation, and is an infinitesimal nontrivial diffeomorphism. The exact all-orders corrected solution is then just the finite diffeomorphism. This obviously is a solution of chiral gravity with no log modes excited.

At the quantum level, the question is trickier. Of course we do not know whether or not either theory exists quantum mechanically. If log gravity does exists as a logarithmic conformal field theory we know it contains chiral gravity as a superselection

[14] If present, terms in $h_{-\rho}$ proportional to $\rho e^{-2\rho}$ may then, as in [9] be eliminated by a trivial diffeomorphism along $\xi = -2e^{-2\rho}h_{-\rho}\partial_+$.

sector. In perturbation theory, chiral gravity has only the massless gravitons which are dual to the quantum stress tensor current algebra. The OPEs of these fields obviously close and hence decouple from log gravity. But we do not know if that superselection sector contains more than just the current algebra, or if it is local or modular invariant. Equivalently, we do not know if the black hole microstates are *exactly* chiral at the quantum level, or contain small nonchiral charges which are lost in the semiclassical limit. As we shall now see, this is more or less equivalent to the question of whether or not there are extremal CFTs with large central charge. More discussion of this point can be found in the concluding section.

7 Quantum partition function

At this point we have seen that with Brown-Henneaux boundary conditions the linearized spectrum of perturbations of chiral gravity around an AdS_3 background includes only right moving boundary gravitons. We will now use this observation to compute the partition function of the quantum theory assuming applicability of the standard Euclidean methodology. Quite nontrivially, we will find that the resulting partition function has a consistent quantum mechanical interpretation. This can be regarded as evidence that the quantum theory actually exists.

We wish to compute the torus partition function, which is defined as the generating function

$$Z(\tau) = Tr\ q^{L_0}\bar{q}^{\bar{L}_0}, \qquad q = e^{2\pi i\tau} \tag{7.1}$$

encoding the spectrum of the theory. $Z(\tau)$ may be thought of as a canonical ensemble partition function at fixed temperature β^{-1} and angular potential $i\theta$ given by $\tau = \frac{1}{2\pi}(\theta + i\beta)$. We will not attempt to perform the trace in (7.1) directly; this would require an understanding of the Hilbert space of chiral gravity at the non-perturbative level. We instead adopt an indirect approach, following [31, 52, 53, 54].

The standard strategy for computing canonical ensemble partitions in quantum field theory is to go to Euclidean signature. In this case $Z(\tau)$, originally formulated in terms of the Hilbert space of the Lorentzian theory, takes the form of a path integral

$$Z(\tau) \sim \int Dg e^{-kI[g]} \tag{7.2}$$

The dependence of this path integral on τ enters through the boundary conditions imposed on the metric g; the boundary is taken to be a two torus T^2 with conformal structure parameter τ. For clarity, we have extracted from equation (7.2) an explicit

factor of the dimensionless coupling constant of the theory, $k = \ell/16G$. In terms of the central charge of the dual boundary theory, $k = c/24$.

At large k the dominant contribution to the path integral is given by the saddle point approximation

$$Z(\tau) = \sum_{g_c} e^{-kI[g_c]+I^{(1)}[g_c]+\frac{1}{k}I^{(2)}[g_c]+\ldots} \tag{7.3}$$

Here the sum is over classical solutions g_c to the equations of motion of the theory. $I[g_c]$ denotes the corresponding classical action. The subleading terms of the form $k^{1-n}I^{(n)}[g_c]$ represent quantum corrections to the effective action at n^{th} order in perturbation theory.

We will take equation (7.3) to be our working definition of the path integral of chiral gravity (7.2) and assume the equivalence of (7.2) with (7.1). In quantum mechanics, this equivalence can be rigorously established. In quantum field theory in general it cannot be proven, but has worked well in many situations. In quantum gravity, the Euclidean approach is less well-founded because, among other problems, the action is unbounded below. Nevertheless, straightforward applications in quantum gravity have tended to yield sensible answers. We will simply assume that this is the case for the path integral of chiral gravity. At the end of this paper we will discuss various ways in which this assumption might fail.

7.1 Classical saddle points

Our first task is to determine which classical saddle points contribute to the partition function (7.3). These saddles g_c are solutions to the classical equations of motion with T^2 conformal boundary. In Euclidean signature, the bulk action of chiral gravity is

$$I_{TMG} = \frac{1}{16\pi G}\int d^3x\sqrt{g}\left(R+\frac{2}{\ell^2}\right) + i\ell\int d^3x\sqrt{g}\varepsilon^{\lambda\mu\nu}\Gamma^r_{\lambda\sigma}\left(\frac{1}{2}\partial_\mu\Gamma^\sigma_{r\nu}+\frac{1}{3}\Gamma^\sigma_{\mu\tau}\Gamma^\tau_{\nu r}\right) \tag{7.4}$$

The factor of i appearing in front of the final term term is the usual one that appears for Chern-Simons theories in Euclidean signature. To see that it must be there, note that the Chern-Simons Lagrangian is a pseudo-scalar rather than a scalar. Hence in Lorentzian signature the Chern-Simons term is odd under time reversal $t \to -t$. Rewriting in terms of the Euclidean time variable $t_E = it$ we see that the Chern-Simons action is pure imaginary in Euclidean signature. The corresponding Euclidean

equations of motion are

$$\mathcal{G}_{\mu\nu} + i\ell C_{\mu\nu} = 0 \tag{7.5}$$

where $\mathcal{G}_{\mu\nu}$ and $C_{\mu\nu}$ are defined as in (2.4). One can verify directly that this is just the Lorenztian equation of motion (2.3) written in terms of a Euclidean time coordinate $t_E = it$.

The classical saddle points are smooth, real[15] Euclidean metrics which solve (7.5). For these metrics both $\mathcal{G}_{\mu\nu}$ and $C_{\mu\nu}$ are real, so must vanish separately. Thus these saddle points obey the equations of motion of Einstein gravity with a negative cosmological constant

$$\mathcal{G}_{\mu\nu} = 0 \tag{7.6}$$

The fact that Euclidean saddle points must be locally Einstein is in contrast with the quite difficult problem of solving the equations of motion Lorentzian signature. This dramatic simplification will allow us to compute the partition function exactly. One might interpret this simplification either as evidence that the Euclidean formulation does not correctly capture the complexity of the Lorentzian theory, or as evidence that the Lorentzian theory has a hidden simplicity. Indeed it is possible that all Lorentzian solutions of chiral gravity are locally Einstein.

Solutions of the equation of motion (7.6) are locally isometric to three dimensional hyperbolic space $\mathbb{H}_3$ with Ricci curvature $R = -6/\ell^2$. So we just need to classify locally hyperbolic three manifolds with T^2 boundary. Any locally hyperbolic three manifold is a quotient of $\mathbb{H}_3$ by a discrete subgroup of its isometry group $SL(2,\mathbb{C})$. In fact, it is straightforward to show (see e.g. [31]) that any such smooth geometry with a T^2 conformal boundary must be of the form $H^3/\mathbb{Z}$. We will not review this classification in detail, but simply summarize the salient points.

We will take the boundary T^2 to be parameterized by a complex coordinate z, in terms of which the periodicity conditions are

$$z \sim z + 1 \sim z + \tau \tag{7.7}$$

This complex coordinate is related to the usual time and angular coordinates of global AdS_3 by $z = \frac{1}{2\pi}(\phi + it)$.[16] To find a Euclidean geometry whose boundary has this

[15] One might wonder whether complex saddle points should be considered. In Euclidean quantum field theory, one is instructed to include complex saddle points when, for example, momenta are held fixed at the boundary. As we are fixing the boundary metric here there is no obvious reason to include complex saddle points.

[16] In equation (2.5) we used τ and ϕ to denote the time and angular coordinates of global AdS_3. Here we use t and ϕ to avoid confusion with the conformal structure parameter τ.

conformal structure, write $\mathbb{H}_3$ in planar coordinates as

$$\frac{ds^2}{\ell^2} = \frac{dw\bar{d}w + dy^2}{y^2} \tag{7.8}$$

The conformal boundary is at $y = 0$, along with the point $y = \infty$. To obtain $\mathbb{H}_3/\mathbb{Z}$ we will quotient by the identification

$$w \sim qw, \qquad q = e^{2\pi i\tau} \tag{7.9}$$

If we identify $w = e^{2\pi iz}$, then the identifications (7.7) follow. We will call the quotient $\mathbb{H}_3/\mathbb{Z}$ constructed in this way $M_{0,1}$.

Now, the geometry described above is not the only locally hyperbolic manifold with the desired boundary behavior. To see this, note that the geometry (7.8) does not treat the two topologically non-trivial cycles of the boundary T^2 in a democratic manner. In particular, the ϕ (real z) cycle of the boundary torus is contractible in the interior of the geometry (7.8), while the t cycle is not. In fact, for every choice of cycle in the boundary T^2 one can find a quotient dimensional manifold $\mathbb{H}_3/\mathbb{Z}$ which makes this cycle contractible. A topologically nontrivial cycle $ct + d\phi$ in T^2 is labeled by a pair of relatively prime integers (c, d). The associated quotient $\mathbb{H}_3/\mathbb{Z}$ will be denoted $M_{c,d}$. These geometries were dubbed the "$SL(2,\mathbb{Z})$ family of black holes" by [55].

To describe these manifolds, consider the group of modular transformations

$$\gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in SL(2,\mathbb{Z}) \tag{7.10}$$

which act as conformal transformations of the boundary T^2. The cycles t and ϕ transform as a vector $\begin{pmatrix} t \\ \phi \end{pmatrix}$ under $SL(2,\mathbb{Z})$, so the element γ takes $\phi \to c\phi + dt$. Under these transformations the conformal structure of the boundary T^2 is invariant, and τ transforms in the usual way

$$\tau \to \gamma\tau = \frac{a\tau + b}{c\tau + d} \tag{7.11}$$

These conformal transformations of the boundary T^2 extend to isometries in the interior. These isometries are easiest to write down by combining the bulk coordinates (w, y) into a single quaternionic coordinate $h = w + jy$. The modular transformation

acts as

$$h \to \gamma h = (ah + b)(ch + d)^{-1} \tag{7.12}$$

Applying this isometry to the geometry $M_{0,1}$ described in (7.8) gives a geometry $M_{c,d}$ in which the cycle $c\phi + dt$ is contractible. This geometry will represent a saddle point contribution to the partition function. Moreover, it is possible to demonstrate the $M_{c,d}$ so constructed are in fact the only smooth real saddle point contributions to the partition sum.

We should emphasize that the pair of relatively prime integers (c, d) determines the geometry $M_{c,d}$ uniquely. Note that (c, d) does not determine $\gamma = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$ uniquely, as a and b are determined only up to an overall shift $(a, b) \to (a + nc, b + nd)$ for some $n \in \mathbb{Z}$. However, one can check that the geometry $M_{c,d}$ is in fact independent of n up to a diffeomorphism which leaves the boundary invariant. Thus the geometries $M_{c,d}$ are in one to one correspondence with elements of the coset $SL(2, \mathbb{Z})/\mathbb{Z}$.

We conclude that the partition function takes the form

$$Z(\tau) = \sum_{(c,d)} Z_{c,d}(\tau) \tag{7.13}$$

where $Z_{c,d}(\tau)$ denotes the contribution from the saddle $M_{c,d}$. Since the geometries are related by modular transformations we may write this as

$$Z(\tau) = \sum_{(c,d)} Z_{0,1}(\gamma\tau) \tag{7.14}$$

where γ is given by (7.10). The sum over (c, d) may be thought of as a sum over the coset $SL(2, \mathbb{Z})/\mathbb{Z}$. Such sums are known as Poincaré series and first appeared in the context of three dimensional gravity in [52].

7.2 Sum over geometries

We now wish to compute the perturbative partition function $Z_{0,1}(\tau)$ around the saddle point geometry $M_{0,1}$ given in (7.8). The computation of the classical piece, including the gravitational Chern-Simons term, was given in [56]. This computation is rather subtle as the appropriate boundary terms must be included in the action. The result is

$$e^{-I[M_{0,1}]} = q^{-k} \tag{7.15}$$

We note that this answer is complex, since the original Euclidean action (7.4) was complex. In particular, equation (7.15) is holomorphic in τ. It is crucial that we are at the chiral point $\mu\ell = 1$, otherwise the action would not be holomorphic in τ.

In order to determine the perturbative corrections to this saddle point action, we will follow the argument of [31]. The geometry $M_{0,1}$ is simply the Euclidean geometry found by imposing the identifications (7.7) on the global t and ϕ coordinates of AdS_3. It is therefore the usual Euclidean geometry associated with the canonical ensemble partition function computed in a fixed Anti-de Sitter background. The partition function $Z_{0,1}$ therefore has the interpretation in Lorenztian signature as

$$Z_{0,1} = Tr_{\mathcal{H}} q^{L_0} \bar{q}^{\bar{L}_0} \tag{7.16}$$

where the trace is over the Hilbert space $\mathcal{H}$ of quantum fluctuations around a fixed Anti-de Sitter background. The classical contribution (7.15) may be interpreted as the contribution to this trace from a ground state $|0\rangle$ of conformal dimension $L_0|0\rangle = -k|0\rangle$. This ground state is just empty Anti-de Sitter space in the absence of any excitations.

At the linearized level, as shown above, the Hilbert space $\mathcal{H}$ includes only right moving boundary gravitons. The linearized metrics of these boundary gravitons are obtained by acting with a right moving Virasoro generator on the empty AdS_3 vacuum state. The generator L_{-1} annihilates the vacuum, as L_{-1} is an isometry of AdS_3. The other L_{-n}, with $n \geq 2$ act as creation operators, and describe non-trivial boundary graviton states. In the CFT language, such a boundary graviton is thought of as a state of the form

$$L_{-i_1} \dots L_{-i_n}|0\rangle, \qquad i_n \geq 2 \tag{7.17}$$

The resulting trace over these states is easy to compute. It is a character of the Virasoro algebra

$$Z_{1,0} = q^{-k} \prod_{n=2}^{\infty} \frac{1}{1-q^n} \tag{7.18}$$

which is closely related to the Dedekind eta function.

It is illustrative to compare this formula to equation (7.3). The trace over Virasoro descendants can be interpreted as the one loop contribution to the free energy; this is to be expected, as the boundary gravitons are solutions to the linearized equations of motion. It would be interesting to derive this result directly by computing an appropriate one loop determinant, as in [57].

We now ask to what extent the formula (7.18) may be altered by higher order corrections in powers of the inverse coupling k^{-1}, i.e. by the terms $I^{(n)}[g_c]$ for $n \geq 2$ in (7.3). We first note that the dimensions of the states appearing in the representation (7.17) are completely fixed by the Virasoro algebra. Once the dimension of the vacuum state is known, the result (7.18) is the only answer consistent with the existence of a Virasoro algebra. This implies that equation (7.18) is one-loop exact, in the sense that the energy levels of the known states can not be altered in perturbation theory. The only possible modification of this formula at higher orders in perturbation theory in $1/k$ is a shift in the dimension of the vacuum state. This shift is interpreted as a renormalization of the cosmological constant. It may be absorbed by a shift of the bare coupling constant order by order in perturbation theory.

One might wonder whether there are additional states which are not present at linear order which might contribute to the sum. We do not claim to have a complete understanding of the Lorentzian spectrum at the non-linear level, and so can not rule out this possibility. If such states do exist, they are not solutions of the Einstein equation and hence do not appear in the Euclidean formulation followed here. So, if the Euclidean methodology assumed here is correct, it implies that either no such additional states/corrections appear or they cancel among themselves. In this case equation (7.18) includes the contributions to the saddle point action to all orders in the perturbation expansion in $1/k$.

Putting together the results of the previous analyses, we conclude that the partition function of chiral gravity takes the form

$$Z(\tau) = \sum_{c,d} Z_{0,1}(\gamma\tau), \qquad Z_{0,1}(\tau) = q^{-k} \prod_{n=2}^{\infty} \frac{1}{1-q^n} \tag{7.19}$$

This sum is naively divergent, but has a well defined regularization (analogous to zeta function regularization) which is consistent with modular invariance. We will not review the details of this regularization, which has been discussed elsewhere [31, 52, 53, 54, 58], but simply state the result.

To start, we expand $Z_{0,1}(\tau)$ in powers of q

$$Z_{0,1}(\tau) = \sum_{\Delta=-k}^{\infty} a(\Delta) q^{\Delta}, \qquad a(\Delta) = p(\Delta+k) - p(\Delta+k-1) \tag{7.20}$$

where $p(N)$ is the number of partitions of the integer N. Then the regularization of

the sum (7.19) is

$$Z(\tau) = \sum_{\Delta'=-k}^{0} a(\Delta') T_{-\Delta'} J \tag{7.21}$$

where $T_n J$ denotes the action of the nth Hecke operator on the modular function $J(\tau)$. From a practical point of view, $T_n J$ may be defined as the unique holomorphic, modular invariant function on the upper half plane which has a pole of order n at $\tau = i\infty$. In particular, it is the unique $SL(2,\mathbb{Z})$ invariant function whose Taylor expansion in integer powers of q is

$$T_n J(\tau) = q^{-n} + \mathcal{O}(q) \tag{7.22}$$

The coefficients in this Taylor expansion are positive integers which are straightforward to compute; we refer the reader to [59] for a more detailed discussion of these Hecke operators and their properties.

7.3 Physical interpretation

The above analysis implies that, with the assumptions noted above, the partition function of chiral gravity takes the form

$$Z(\tau) = \sum_{\Delta=-k}^{\infty} N(\Delta) q^{\Delta} \tag{7.23}$$

where the $N(\Delta)$ are positive integers. These positive integers may be computed for any desired value of k, as described in detail in [32]. In fact, the partition function (7.23) is precisely the holomorphic part of the partition function conjectured to be dual to pure gravity in [32]. This is not a coincidence, as chiral gravity apparently is a theory with all the properties shown in [32] to lead uniquely to (7.23). As this partition function contains as few states as possible consistent with modular invariance, it is referred to as the extremal CFT partition function.

This partition function is exactly of the form that one expects for a consistent quantum theory; it is a discrete sum over a positive spectrum, with positive integer coefficients. We contrast the present situation with that of pure Einstein gravity [31], where the corresponding computation did not yield a consistent quantum mechanical partition function unless complexified geometries were included in the sum. The inclusion of the gravitational Chern-Simons term has resolved this apparent inconsistency.

The partition function (7.23) has several additional interesting properties. First, we note that the partition function makes sense only when k is an integer. Thus the cosmological constant and the Chern-Simons coefficient are quantized in Planck units. Moreover, the spectrum of dimensions is quantized, $\Delta \in \mathbb{Z}$. Thus the masses and angular momenta of all states in the theory – including black holes – are quantized as well.

These two rather remarkable statements are consequences of the fact that the theory is chiral. To see this, note that in a chiral theory the partition function $Z(\tau)$ must depend holomorphically on τ. The complex structure of the boundary T^2 is modular invariant, so we may think of $Z(\tau)$ as a holomorphic function on the modular domain $\mathbb{H}_2/SL(2,\mathbb{Z})$. Including the point at $\tau = i\infty$ this modular domain may be thought of as a Riemann surface of genus zero, which is mapped analytically to the usual Riemann sphere $\mathbb{C} \cup \{\infty\}$ by the j-invariant $J(\tau)$ (see e.g. [59]). Since the partition function $Z(\tau)$ is meromorphic, it is therefore a rational function of $J(\tau)$. Moreover, if we assume that the canonical ensemble partition sum is convergent, $Z(\tau)$ must be holomorphic at all points, except possibly at $\tau = i\infty$. Thus $Z(\tau)$ is a polynomial in the J-invariant

$$Z = \sum_{n \geq 0} a_n J(\tau)^n, \qquad J(\tau) = q^{-1} + 744 + 196884q + \ldots \tag{7.24}$$

for some real coefficients a_n It follows that both the coupling constant and the spectrum of dimensions are quantized.

We should note, however, that this argument does not imply that the coefficients $N(\Delta)$ appearing in the expansion are positive integers. This fact was crucial for a consistent quantum mechanical interpretation of the partition function.

These coefficients $N(\Delta)$ for large Δ can be interpreted as the exact degeneracies of quantum black holes in chiral gravity. One can demonstrate the these coefficients reproduce precisely the black hole entropy, including an infinite series of corrections. This is done by reorganizing the modular sum (7.19) into a Rademacher expansion [31, 52, 58]. The computation proceeds exactly as in [31], so we refer the reader there for details.

Finally, we emphasize that it is not at all clear that conformal field theories with the spectrum described above exist. No examples have been constructed with $k > 1$. Indeed, a potential objection to the existence of these theories at large k was noted in [60, 61].[17] Although the results of this paper do not imply the existence of such

[17] See also [62] for a discussion of these objections and [63] for a related discussion in the context

extremal CFTs, they certainly fit harmoniously with their conjectured existence.

7.4 What could go wrong

We have argued that (7.23) follows from a conservative set of assumptions. Nevertheless, our argument is not watertight. We now list some possible reasons why (7.23) might not actually be the quantum chiral gravity partition function.

- The Euclidean approach is invalid because the path integral is unbounded.
- Other complex saddle points are encountered in the analytic continuation from Lorentzian to Euclidean signature and must be included.
- There are nonperturbative Lorentzian classical solutions other than black holes which correspond to additional primaries in the CFT and non-perturbative corrections to the Euclidean saddle point approximation.
- Non-smooth saddle points must be included.

These various possibilities are not mutually exclusive.

8 Chiral gravity, log gravity, extremal CFT and log extremal CFT

We have presented several results pertaining to chiral gravity, log gravity, extremal CFT and their interconnections. Much remains to be understood. In this concluding section we will draw lessons from what we learned and freely speculate on various possible outcomes. There are many possibilities – we will limit ourselves below to the most pessimistic and the two most optimistic outcomes.

8.1 Nothing makes sense

The least interesting possibility, which cannot be excluded, is that none of the theories under discussion are physically sensible. It might turn out that at the classical level chiral gravity has negative energy solutions, non-perturbative instabilities and/or generically develops naked singularities. In this case the quantum theory is unlikely to be well defined. If chiral gravity is not classically sensible, log gravity – which contains chiral gravity – is not likely to be well defined either. Extremal CFTs

of supersymmetric theories.

with large central charges may simply not exist. The main contraindicator to this possibility is that we have so far discovered a rich and cohesive mathematical structure with no apparent internal inconsistencies.

8.2 Chiral gravity = extremal CFT

An obvious and interesting possibility is that chiral gravity is fully consistent and unitary, and has the modular invariant partition function proposed in [32]. In this case chiral gravity is holographically dual to a local extremal CFT . There seems to be no room here for non-Einstein Lorentzian solutions of chiral gravity because there are no corresponding primaries in an extremal CFT. An important indicator in favor of this scenario is that the torus partition function, formally computed using Euclidean methods, gives the extremal CFT partition function on the nose. In this case the genus g partition function of the extremal CFT would be simply the chiral-gravity weighted sum over geometries with genus g boundary. Conversely, if the extremal CFTs are constructed, we are finished: we can simply declare them, in the spirit of [32], to be the quantization of chiral gravity.

Of course extremal CFTs have not been constructed for $k > 1$. Indeed, arguments against the existence of extremal CFTs at large k were described in [60, 61], although no proof was given. An optimist might view the failure of these valiant efforts to produce an actual non-existence proof as indication that extremal CFTs do exist as highly exceptional mathematical objects. A pessimist, on other other hand, might take the fact that no extremal CFTs have been constructed for $k > 1$ as evidence that they do not exist. Further investigation is clearly needed.

8.3 Chiral gravity $\in$ log gravity=log extremal CFT

A third interesting possibility is the following.[18] Assume that quantum log gravity exists and has a well-defined Hilbert space, and that there is a holographically dual CFT which is logarithmic and not chiral. Of course, this is not of so much interest in and of itself, as there is no shortage of non-unitary quantum theories of gravity. However, chiral gravity then also necessarily exists as the superselection sector in which all left charges vanish. This superselection sector could still itself have undesirable properties. In particular there is no a priori guarantee that it is modular invariant. Since a modular transformation is a large diffeomorphism in Euclidean space, this is certainly a desirable property. Modular invariance should be violated if

[18] We are grateful to V. Schomerus for discussions on this point.

the chiral states are in some sense incomplete. For example, consider the truncation of a generic non-chiral CFT to the purely right-chiral sector. Generically, the only chiral operators are the descendants of the identity created with the right moving stress-tensor. The partition function is simply a Virasoro character and is not modular invariant. In the previous section it was argued that in the context of chiral gravity the primaries associated to black holes complete, in the manner described in [32], this character to a modular invariant partition function. However, it is possible that no such completion exists. It might be that chiral gravity is in some sense a physical, unitary subsector of log gravity, but its dual does not obey all the axioms of a local CFT. Interestingly, in [49] it was found that some 2D gravity theories coupled to matter are logarithmic CFTs.

At first the compelling observation that the Euclidean computation of the chiral gravity partition function gives the extremal CFT partition function would seem to be evidence against this possibility. One would expect that any extra states present in log gravity would spoil this nice result. However, as log gravity is not unitary, the extra contributions to the partition function can vanish or cancel. Indeed it is a common occurrence in logarithmic CFTs for the torus partition function to contain no contributions from the logarithmic partners. We see hints of this here: as $c_L = 0$, the left-moving gravitons of log gravity have zero norm and hence do not contribute. This suggests the at-present-imprecise notion of a "log extremal CFT": a logarithmic CFT whose partition function is precisely the known extremal partition function. Perhaps previous attempts to construct extremal CFTs have failed precisely because the theory was assumed to be unitary rather than logarithmic. Clearly there is much to be understood and many interesting avenues to pursue.

Acknowledgements

This work was partially funded by DOE grant DE-FG02-91ER40654. The work of W. S. was also funded by NSFC. The work of A. M. is supported by the National Science and Engineering Research Council of Canada. We wish to thank S. Carlip, S. Deser, T. Hartman, M. Gaberdiel, S. Giombi, G. Giribet, D. Grumiller, C. Keller, M. Kleban, G. Moore, M. Porrati, O. Saremi, V. Schomerus, E. Witten, A. Wissanji and X. Yin for useful conversations and correspondence. W. S. thanks the High Energy Group at Harvard and McGill Physics Department for their kind hospitality.

References

[1] W. Li, W. Song and A. Strominger, "Chiral Gravity in Three Dimensions," JHEP **0804**, 082 (2008) [arXiv:0801.4566 [hep-th]].

[2] S. Deser, R. Jackiw and S. Templeton, "Topologically massive gauge theories," Annals Phys. **140**, 372 (1982) [Erratum-ibid. **185**, 406.1988 APNYA,281,409 (1988 APNYA,281,409-449.2000)].

[3] S. Deser, R. Jackiw and S. Templeton, "Three-Dimensional Massive Gauge Theories," Phys. Rev. Lett. **48**, 975 (1982).

[4] C. Fefferman and C.R. Graham, "Conformal Invariants, in *Elie Cartan et les Mathematiques daujourdhui* (Asterisque, 1985) 95.

[5] J. D. Brown and M. Henneaux, "Central Charges in the Canonical Realization of Asymptotic Symmetries: An Example from Three-Dimensional Gravity," Commun. Math. Phys. **104**, 207 (1986).

[6] S. Carlip, S. Deser, A. Waldron and D. K. Wise, "Cosmological Topologically Massive Gravitons and Photons," arXiv:0803.3998v1 [hep-th].

[7] S. Carlip, S. Deser, A. Waldron and D. K. Wise, "Cosmological Topologically Massive Gravitons and Photons," Class. Quant. Grav. **26**, 075008 (2009) [arXiv:0803.3998v3 [hep-th]].

[8] G. Giribet, M. Kleban and M. Porrati, "Topologically Massive Gravity at the Chiral Point is Not Chiral," arXiv:0807.4703v1 [hep-th].

[9] G. Giribet, M. Kleban and M. Porrati, "Topologically Massive Gravity at the Chiral Point is Not Unitary," arXiv:0807.4703v2 [hep-th].

[10] D. Grumiller and N. Johansson, "Instability in cosmological topologically massive gravity at the chiral point," arXiv:0805.2610 [hep-th].

[11] M. i. Park, "Constraint Dynamics and Gravitons in Three Dimensions," arXiv:0805.4328 [hep-th].

[12] D. Grumiller, R. Jackiw and N. Johansson, "Canonical analysis of cosmological topologically massive gravity at the chiral point," arXiv:0806.4185 [hep-th].

[13] S. Carlip, S. Deser, A. Waldron and D. K. Wise, "Topologically Massive AdS Gravity," arXiv:0807.0486 [hep-th].

[14] S. Carlip, "The Constraint Algebra of Topologically Massive AdS Gravity," arXiv:0807.4152v1 [hep-th].

[15] A. Strominger, "A Simple Proof of the Chiral Gravity Conjecture," arXiv:0808.0506 [hep-th], to appear proceedings of the Symposium in Honor of S. T. Yau.

[16] S. Carlip, "The Constraint Algebra of Topologically Massive AdS Gravity," arXiv:0807.4152v3 [hep-th].

[17] W. Li, W. Song and A. Strominger, "Comment on 'Cosmological Topological Massive Gravitons and Photons'," arXiv:0805.3101 [hep-th].

[18] I. Sachs and S. N. Solodukhin, "Quasi-Normal Modes in Topologically Massive Gravity," JHEP **0808**, 003 (2008) [arXiv:0806.1788 [hep-th]].

[19] H. W. Lee, Y. W. Kim and Y. S. Myung, "Quasinormal modes for topologically massive black hole," arXiv:0807.1371 [hep-th].

[20] I. Sachs, "Quasi-Normal Modes for Logarithmic Conformal Field Theory," JHEP **0809**, 073 (2008) [arXiv:0807.1844 [hep-th]].

[21] G. W. Gibbons, C. N. Pope and E. Sezgin, "The General Supersymmetric Solution of Topologically Massive Supergravity," Class. Quant. Grav. **25**, 205005 (2008) [arXiv:0807.2613 [hep-th]].

[22] Y. S. Myung, "Logarithmic conformal field theory approach to topologically massive gravity," Phys. Lett. B **670**, 220 (2008) [arXiv:0808.1942 [hep-th]].

[23] M. Alishahiha, R. Fareghbal and A. E. Mosaffa, "2D Gravity on AdS_2 with Chern-Simons Corrections," JHEP **0901**, 069 (2009) [arXiv:0812.0453 [hep-th]].

[24] W. Kim and E. J. Son, "Chiral black hole in three-dimensional gravitational Chern-Simons," arXiv:0812.4790 [hep-th].

[25] E. Sezgin and Y. Tanii, "Witten-Nester Energy in Topologically Massive Gravity," arXiv:0903.3779 [hep-th].

[26] Y. B. Choquet and S. Deser "Stability of Flat Space" Ann. Phys. **81**, 165 (1973).

[27] Geroch, Robert, "Is perturbation theory misleading in general relativity?" J. Math. Phys. **26**, 10 (1985).

[28] D. Brill and C. V. Vishveshwara, "Joint linearization instabilities in general realtivity" J. Math. Phys. **27**, 1813 (1986).

[29] M. Henneaux, C. Martinez and R. Troncoso, "Asymptotically anti-de Sitter spacetimes in topologically massive gravity," arXiv:0901.2874 [hep-th].

[30] M. Blagojevic and B. Cvetkovic, "Canonical structure of topologically massive gravity with a cosmological constant," arXiv:0812.4742 [gr-qc].

[31] A. Maloney and E. Witten, "Quantum Gravity Partition Functions in Three Dimensions," arXiv:0712.0155 [hep-th].

[32] E. Witten, "Three-Dimensional Gravity Revisited," arXiv:0706.3359 [hep-th].

[33] D. Grumiller and N. Johansson, "Consistent boundary conditions for cosmological topologically massive gravity at the chiral point," arXiv:0808.2575 [hep-th].

[34] E. Ayon-Beato and M. Hassaine, Annals Phys. **317**, 175 (2005) [arXiv:hep-th/0409150].

[35] A. Garbarz, G. Giribet and Y. Vasquez, "Asymptotically AdS_3 Solutions to Topologically Massive Gravity at Special Values of the Coupling Constants," arXiv:0811.4464 [hep-th].

[36] K. Hotta, Y. Hyakutake, T. Kubota and H. Tanida, "Brown-Henneaux's Canonical Approach to Topologically Massive Gravity," JHEP **0807**, 066 (2008) [arXiv:0805.2005 [hep-th]].

[37] G. Barnich and F. Brandt, "Covariant theory of asymptotic symmetries, conservation laws and central charges," Nucl. Phys. B **633**, 3 (2002) [arXiv:hep-th/0111246].

[38] G. Barnich and G. Compere, "Surface charge algebra in gauge theories and thermodynamic integrability," J. Math. Phys. **49**, 042901 (2008) [arXiv:0708.2378 [gr-qc]].

[39] L. F. Abbott and S. Deser, "Stability Of Gravity With A Cosmological Constant," Nucl. Phys. B **195**, 76 (1982).

L. F. Abbott and S. Deser, "Charge Definition In Nonabelian Gauge Theories," Phys. Lett. B **116**, 259 (1982).

V. Iyer and R. M. Wald, "Some properties of Noether charge and a proposal for dynamical black hole entropy," Phys. Rev. D **50**, 846 (1994) [arXiv:gr-qc/9403028].

G. Barnich, F. Brandt and M. Henneaux, "Local Brst Cohomology In The Antifield Formalism. 1. General Theorems," Commun. Math. Phys. **174**, 57 (1995) [arXiv:hep-th/9405109].

I. M. Anderson and C. G. Torre, "Asymptotic conservation laws in field theory," Phys. Rev. Lett. **77**, 4109 (1996) [arXiv:hep-th/9608008].

C. G. Torre, "Local cohomology in field theory with applications to the Einstein equations," arXiv:hep-th/9706092.

G. Barnich, F. Brandt and M. Henneaux, "Local BRST cohomology in gauge theories," Phys. Rept. **338**, 439 (2000) [arXiv:hep-th/0002245].

[40] G. Barnich, "Boundary charges in gauge theories: Using Stokes theorem in the bulk," Class. Quant. Grav. **20**, 3685 (2003) [arXiv:hep-th/0301039].

G. Barnich, F. Brandt and K. Claes, "Asymptotically anti-de Sitter space-times: symmetries and conservation laws revisited," Nucl. Phys. Proc. Suppl. **127**, 114 (2004) [arXiv:gr-qc/0306112].

G. Barnich and G. Compere, "Conserved charges and thermodynamics of the spinning Goedel black hole," Phys. Rev. Lett. **95**, 031302 (2005) [arXiv:hep-th/0501102].

M. Banados, G. Barnich, G. Compere and A. Gomberoff, Phys. Rev. D **73**, 044006 (2006) [arXiv:hep-th/0512105].

G. Compere, "Symmetries and conservation laws in Lagrangian gauge theories with applications to the mechanics of black holes and to gravity in three dimensions," arXiv:0708.3153 [hep-th].

[41] J. D. Brown and J. W. . York, "Quasilocal energy and conserved charges derived from the gravitational action," Phys. Rev. D **47**, 1407 (1993) [arXiv:gr-qc/9209012].

[42] S. Deser and B. Tekin, "Energy in topologically massive gravity," Class. Quant. Grav. **20**, L259 (2003) [arXiv:gr-qc/0307073].

[43] A. Bouchareb and G. Clement, "Black hole mass and angular momentum in topologically massive gravity," Class. Quant. Grav. **24**, 5581 (2007) [arXiv:0706.0263 [gr-qc]].

[44] G. Compere and S. Detournay, "Semi-classical central charge in topologically massive gravity," arXiv:0808.1911 [hep-th].

[45] G. Compere and S. Detournay, "Centrally extended symmetry algebra of asymptotically Goedel spacetimes," JHEP **0703**, 098 (2007) [arXiv:hep-th/0701039].

[46] A. N. Aliev and Y. Nutku, Class. Quant. Grav. **13**, L29 (1996) [arXiv:gr-qc/9812089].

[47] M. Cavaglia, Grav. Cosmol. **5**, 101 (1999) [arXiv:gr-qc/9904047].

[48] K. A. Moussa, G. Clement and C. Leygnac, "The black holes of topologically massive gravity," Class. Quant. Grav. **20**, L277 (2003) [arXiv:gr-qc/0303042].

[49] A. Bilal and I. I. Kogan, "On gravitational dressing of 2-D field theories in chiral gauge," Nucl. Phys. B **449**, 569 (1995) [arXiv:hep-th/9503209].

[50] V. Gurarie, "Logarithmic operators in conformal field theory," Nucl. Phys. B **410**, 535 (1993) [arXiv:hep-th/9303160].

[51] C. Crnkovic and E. Witten, "Covariant Description Of Canonical Formalism In Geometrical Theories", in S. W. Hawking and W. Israel : Three hundred years of gravitation*, p.676-684.

A. M. Ghezelbash, M. Khorrami and A. Aghamohammadi, "Logarithmic conformal field theories and AdS correspondence," Int. J. Mod. Phys. A **14**, 2581 (1999) [arXiv:hep-th/9807034].

I. I. Kogan, "Singletons and logarithmic CFT in AdS/CFT correspondence," Phys. Lett. B **458**, 66 (1999) [arXiv:hep-th/9903162].

Y. S. Myung and H. W. Lee, "Gauge bosons and the AdS(3)/LCFT(2) correspondence," JHEP **9910**, 009 (1999) [arXiv:hep-th/9904056].

A. Lewis, "Logarithmic operators in AdS(3)/CFT(2)," Phys. Lett. B **480**, 348 (2000) [arXiv:hep-th/9911163].

M. Flohr, "Bits and pieces in logarithmic conformal field theory," Int. J. Mod. Phys. A **18**, 4497 (2003) [arXiv:hep-th/0111228].

M. R. Gaberdiel, "An algebraic approach to logarithmic conformal field theory," Int. J. Mod. Phys. A **18**, 4593 (2003) [arXiv:hep-th/0111260].

S. Moghimi-Araghi, S. Rouhani and M. Saadat, "Correlation functions and AdS/LCFT correspondence," Nucl. Phys. B **696**, 492 (2004) [arXiv:hep-th/0403150].

M. Henneaux, C. Martinez, R. Troncoso and J. Zanelli, "Asymptotically anti-de Sitter spacetimes and scalar fields with a logarithmic branch," Phys. Rev. D **70**, 044034 (2004) [arXiv:hep-th/0404236].

M. Henneaux, C. Martinez, R. Troncoso and J. Zanelli, "Asymptotic behavior and Hamiltonian analysis of anti-de Sitter gravity coupled to scalar fields," Annals Phys. **322**, 824 (2007) [arXiv:hep-th/0603185].

[52] R. Dijkgraaf, J. M. Maldacena, G. W. Moore and E. P. Verlinde, "A black hole farey tail," arXiv:hep-th/0005003.

[53] J. Manschot, "AdS_3 Partition Functions Reconstructed," JHEP **0710**, 103 (2007) [arXiv:0707.1159 [hep-th]].

[54] X. Yin, "Partition Functions of Three-Dimensional Pure Gravity," arXiv:0710.2129 [hep-th].

[55] J. M. Maldacena and A. Strominger, "AdS(3) black holes and a stringy exclusion principle," JHEP **9812**, 005 (1998) [arXiv:hep-th/9804085].

[56] P. Kraus and F. Larsen, "Partition functions and elliptic genera from supergravity," JHEP **0701**, 002 (2007) [arXiv:hep-th/0607138].

[57] S. Giombi, A. Maloney and X. Yin, "One-loop Partition Functions of 3D Gravity," JHEP **0808**, 007 (2008) [arXiv:0804.1773 [hep-th]].

[58] J. Manschot and G. W. Moore, "A Modern Farey Tail," arXiv:0712.0573 [hep-th].

[59] T. Apostol, *Modular Functions And Dirichlet Series In Number Theory*, Springer.

[60] M. R. Gaberdiel, "Constraints on extremal self-dual CFTs," JHEP **0711**, 087 (2007) [arXiv:0707.4073 [hep-th]].

[61] M. R. Gaberdiel and C. A. Keller, "Modular differential equations and null vectors," JHEP **0809**, 079 (2008) [arXiv:0804.0489 [hep-th]].

[62] D. Gaiotto, "Monster symmetry and Extremal CFTs," arXiv:0801.0988 [hep-th].

[63] M. R. Gaberdiel, S. Gukov, C. A. Keller, G. W. Moore and H. Ooguri, "Extremal N=(2,2) 2D Conformal Field Theories and Constraints of Modularity," arXiv:0805.4216 [hep-th].

arXiv:0907.4529v3 [math.RT] 12 Apr 2012

Rademacher sums, moonshine and gravity

John F. R. Duncan*
Igor B. Frenkel†

1 March 2012

Abstract

In 1939 Rademacher derived a conditionally convergent series expression for the elliptic modular invariant, and used this expression—the first Rademacher sum—to verify its modular invariance. By generalizing Rademacher's approach we construct bases for the spaces of automorphic integrals of arbitrary even integer weight, for groups commensurable with the modular group. Our methods provide explicit expressions for the Fourier expansions of the Rademacher sums we construct at arbitrary cusps, and illuminate various aspects of the structure of the spaces of automorphic integrals, including the actions of Hecke operators.

We give a moduli interpretation for a class of groups commensurable with the modular group which includes all those that are associated to the Monster via monstrous moonshine. We show that within this class the monstrous groups can be characterized just in terms of the behavior of their Rademacher sums. In particular, the genus zero property of monstrous moonshine is encoded naturally in the properties of Rademacher sums.

Just as the ellptic modular invariant gives the graded dimension of the moonshine module, the exponential generating function of the Rademacher sums associated to the modular group furnishes the bi-graded dimension of the Verma module for the Monster Lie algebra. This result generalizes naturally to all the groups of monstrous moonshine, and recovers a certain family of monstrous Lie algebras recently introduced by Carnahan.

Our constructions suggest conjectures relating monstrous moonshine to a distinguished family of chiral three dimensional quantum gravities, and relating monstrous Lie algebras and their Verma modules to the second quantization of this family of chiral three dimensional quantum gravities.

*Case Western Reserve University, Department of Mathematics, 10900 Euclid Avenue, Cleveland, OH 44106.
†Yale University, Department of Mathematics, 10 Hillhouse Avenue, New Haven, CT 06520,
The research of I.F. was supported by NSF grant DMS-0457444.

Contents

1 Introduction **3**
1.1 Monstrous moonshine . . . 3
1.2 Rademacher sums . . . 4
1.3 Conjugate Rademacher sums . . . 6
1.4 Solid tori . . . 7
1.5 Moonshine via Rademacher sums . . . 8
1.6 Modified Rademacher sums . . . 9
1.7 Hecke operators . . . 13
1.8 Monstrous Lie algebras . . . 15
1.9 Chiral gravity . . . 16

2 Conventions **18**
2.1 Functions . . . 18
2.2 Divisors . . . 19
2.3 Isometries . . . 19
2.4 Cosets . . . 22
2.5 Groups . . . 23
2.6 Scalings . . . 25
2.7 Integrals . . . 28

3 Rademacher sums **32**
3.1 Construction . . . 32
3.2 Coefficients . . . 36
3.3 Convergence . . . 44
3.4 Variance . . . 49

4 Modified Rademacher sums **56**
4.1 Construction . . . 56
4.2 Coefficients . . . 59
4.3 Dirichlet series . . . 60
4.4 Convergence . . . 61
4.5 Variance . . . 68

5 Structural applications **70**
5.1 Constants . . . 70
5.2 Inner products . . . 71
5.3 Branching . . . 72

5.4 Fractional orders . . . 77
5.5 Hecke operators . . . 81

6 Moonshine **85**
6.1 Genera . . . 86
6.2 Moduli . . . 87
6.3 Exponents . . . 91
6.4 Cusps . . . 92
6.5 Rademacher sums and the Monster . . . 94

7 Gravity **95**
7.1 First conjecture . . . 95
7.2 Monstrous Lie algebras . . . 98
7.3 Second conjecture . . . 103

References **108**

1 Introduction

1.1 Monstrous moonshine

A remarkable coincidence between the coefficients of the elliptic modular invariant

$$J(z) = \mathrm{e}(-z) + \sum_{n>0} c(n)\,\mathrm{e}(nz), \quad \mathrm{e}(nz) = \exp(2\pi \mathbf{i} n z), \tag{1.1.1}$$

and the dimensions of the irreducible representations of the *Monster group*, denoted $\mathbb{M}$, led McKay and Thompson [Tho79b] to conjecture the existence of a naturally defined infinite dimensional representation

$$V = V_{-1} \oplus V_1 \oplus V_2 \oplus \cdots \tag{1.1.2}$$

for the Monster group, with the property that $\dim V_n = c(n)$. Thompson [Tho79a] also proposed to consider the modular properties of the series

$$T_g(z) = \mathrm{e}(-z) + \sum_{n>0} (\mathrm{tr}|_{V_n} g)\,\mathrm{e}(nz), \tag{1.1.3}$$

for any $g \in \mathbb{M}$, the function $J(z)$ of (1.1.1) being recovered upon taking $g = e$ in (1.1.3). It is clear from the presentation (1.1.3) that $T_g(z)$ is invariant under the *translation group*, which we denote $B(\mathbb{Z})$ and which is generated by the translation $z \mapsto z + 1$. In the case (1.1.1) that g is

the identity element, the invariance extends to the full modular group $PSL_2(\mathbb{Z})$. This leads us to the question: what is special about the invariance groups Γ_g of the McKay–Thompson series (1.1.3)?

In their remarkable paper [CN79], Conway and Norton collected an overwhelming number of coincidences and, in particular, formulated the *moonshine conjecture*:

> The series $T_g(z)$ is the normalized hauptmodul of a genus zero group Γ_g lying between $\Gamma_0(N)$ and its normalizer in $PSL_2(\mathbb{R})$, for each $g \in \mathbb{M}$.

The conjecture of McKay and Thompson was proven in [FLM88] by the explicit construction of a vertex operator algebra $V^\natural$ invariant under the Monster group. It was also shown in [FLM88] that the McKay–Thompson series $T_g(z)$ satisfy the moonshine conjecture of Conway and Norton for all elements g in a subgroup of $\mathbb{M}$ arising as an involution centralizer. Finally, a complete proof of the moonshine conjecture was given by Borcherds in [Bor92]. However, in spite of all the subsequent spectacular developments related to the Monster, the consensus of the experts (cf. e.g. [CMS04]), is that "the real nature of moonshine is still remote".

1.2 Rademacher sums

In this paper we shed new light on the properties of the McKay–Thompson series associated to the Monster, including the mysterious genus zero phenomena. It turns out that the McKay–Thompson series admit presentations as sums over the coset space $B(\mathbb{Z})\backslash\Gamma_g$. The prototype for this is the following expression for the modular invariant $J(z)$ which was originally derived by Rademacher in [Rad39].

$$J(z) + 12 = \mathrm{e}(-z) + \lim_{K\to\infty} \sum_{\substack{0<c<K \\ -K^2<d<K^2 \\ (c,d)=1}} \mathrm{e}\left(-\frac{az+b}{cz+d}\right) - \mathrm{e}\left(-\frac{a}{c}\right) \tag{1.2.1}$$

The integers a and b occurring in each summand of (1.2.1) are chosen so that $ad - bc = 1$. This expression (1.2.1) is remarkably simple, but the convergence is rather subtle. (See [Kno90] for a nice exposition.) The subtraction of the constant $\mathrm{e}(-a/c)$ in (1.2.1) ensures the existence of the limit in what is otherwise a highly divergent series. The Rademacher sum (1.2.1) has been generalized to various discrete subgroups of $PSL_2(\mathbb{R})$ in a series of papers by Knopp (cf. [Kno61a], [Kno61b], [Kno62a], [Kno62b]).

Given $\Gamma < PSL_2(\mathbb{R})$ we write Γ_∞ for the subgroup of Γ that fixes ∞. If Γ is commensurable with the modular group $PSL_2(\mathbb{Z})$ and has the property that $\Gamma_\infty = B(\mathbb{Z})$, then we may naturally

associate a Rademacher sum to Γ by setting

$$R_\Gamma^{(-1)}(z) = \mathrm{e}(-z) + \lim_{K\to\infty} \sum_{\gamma\in(B(\mathbb{Z})\backslash\Gamma)^\times_{\leq K}} \mathrm{e}(-\gamma\cdot z) - \mathrm{e}(-\gamma\cdot\infty), \tag{1.2.2}$$

where the sum here is taken over the *rectangle*

$$(B(\mathbb{Z})\backslash\Gamma)^\times_{\leq K} = \left\{ \gamma = B(\mathbb{Z}) \begin{bmatrix} * & * \\ c & d \end{bmatrix} \in B(\mathbb{Z})\backslash\Gamma \mid 0 < c \leq K,\ -K^2 \leq d \leq K^2 \right\} \tag{1.2.3}$$

(cf. §3.1). In (1.2.3) we write $\left[\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right]$ for the image in $PSL_2(\mathbb{R})$ of a matrix $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$ in $SL_2(\mathbb{R})$, and in (1.2.2) we write $\gamma\cdot z$ for $(az+b)/(cz+d)$ and $\gamma\cdot\infty$ for a/c when $\gamma = \left[\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right]$.

In general the series $R_\Gamma^{(-1)}(z)$ defines an abelian integral for Γ, meaning that we have

$$R_\Gamma^{(-1)}(\gamma\cdot z) = R_\Gamma^{(-1)}(z) + \omega(\gamma) \tag{1.2.4}$$

for some function $\omega : \Gamma \to \mathbb{C}$. We show in §6.1 that the Rademacher sum $R_\Gamma^{(-1)}(z)$ is Γ-invariant if and only if Γ has genus zero. Even more than this, we show (also in §6.1) that when Γ has genus zero the function

$$R_\Gamma^{(-1)}(z) - \frac{1}{2}\mathrm{c}_\Gamma(-1,0) \tag{1.2.5}$$

is the normalized hauptmodul for Γ (cf. §2.5), for a certain constant $\mathrm{c}_\Gamma(-1,0)$. This constant $\mathrm{c}_\Gamma(-1,0)$ turns out to be the *Rademacher constant associated to* Γ as defined in [Nor84] in the case that Γ has genus zero (cf. §5.1).

Rademacher's proof of the validity of the presentation (1.2.1) relies upon explicit formulas for the Fourier coefficients of $J(z)$. These formulas may be given in terms of Kloosterman sums and Bessel functions, and are recovered from the expression

$$\mathrm{c}_\Gamma(m,n) = \sum_{\gamma\in B(\mathbb{Z})\backslash\Gamma/B(\mathbb{Z})} \mathrm{Kl}_\gamma(m,n)\,\mathrm{Bl}_\gamma(m,n) \tag{1.2.6}$$

upon taking $\Gamma = PSL_2(\mathbb{Z})$ and $m = -1$, and allowing n to range over $\mathbb{Z}$. Writing $c(\gamma)$ for $|c|$ when $\gamma = \left[\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right]$, the functions $\mathrm{Kl}_\gamma(m,n)$ and $\mathrm{Bl}_\gamma(m,n)$ are defined as follows (cf. §3.2), for $\gamma\in\Gamma$.

$$\mathrm{Kl}_\gamma(m,n) = \mathrm{e}(m\gamma\cdot\infty)\,\mathrm{e}(-n\gamma^{-1}\cdot\infty) \tag{1.2.7}$$

$$\mathrm{Bl}_\gamma(m,n) = 2\pi\mathbf{i}\,\mathrm{Res}_{\xi=0}\,\mathrm{e}(mc(\gamma)^{-2}\xi^{-1})\,\mathrm{e}(n\xi)\mathrm{d}\xi \tag{1.2.8}$$

Note that $\mathrm{Kl}_\gamma(m,n)$ and $\mathrm{Bl}_\gamma(m,n)$ are well-defined complex numbers only when $c(\gamma)\neq 0$, and

in this case

$$\mathrm{Kl}_\gamma(m,n) = \mathrm{e}\left(\frac{ma+nd}{c}\right), \quad \mathrm{Bl}_\gamma(m,n) = \sum_{k\geq 0}\left(\frac{2\pi\mathbf{i}}{c}\right)^{2k+2}\frac{m^{k+1}}{(k+1)!}\frac{n^k}{k!}, \tag{1.2.9}$$

for $\gamma = \left[\begin{smallmatrix} a & b \\ c & d\end{smallmatrix}\right]$ and $c > 0$. The expression (1.2.6) with $\Gamma = PSL_2(\mathbb{Z})$ and $m = -1$ is, up to elementary transformations, the formula for the n-th coefficient of the elliptic modular invariant given originally by Rademacher in [Rad38]. More generally, the formula (1.2.6) recovers the Fourier coefficients (other than the constant term) of the Rademacher sum $R_\Gamma^{(-1)}(z)$, for Γ an arbitrary group commensurable with $PSL_2(\mathbb{Z})$ and satisfying $\Gamma_\infty = B(\mathbb{Z})$. The following formula (cf. Theorem 3.3.2) encodes the relationship precisely.

$$R_\Gamma^{(-1)}(z) = \mathrm{e}(-z) + \frac{1}{2}\mathrm{c}_\Gamma(-1,0) + \sum_{n>0}\mathrm{c}_\Gamma(-1,n)\,\mathrm{e}(nz) \tag{1.2.10}$$

1.3 Conjugate Rademacher sums

It is striking that the constant term in the Fourier expansion (1.2.10) of the Rademacher sum $R_\Gamma^{(-1)}(z)$ is $\mathrm{c}_\Gamma(-1,0)/2$ and not $\mathrm{c}_\Gamma(-1,0)$, given that the coefficient of $\mathrm{e}(nz)$ in (1.2.10) is exactly $\mathrm{c}_\Gamma(-1,n)$ for $n > 0$. One may also observe that the formula (1.2.6) defining the values $\mathrm{c}_\Gamma(-1,n)$ makes sense for arbitrary integers $n \in \mathbb{Z}$, and so it is natural to ask what rôle is played by the $\mathrm{c}_\Gamma(-1,n)$ for $n < 0$? An answer to this question is obtained when we consider the *conjugate Rademacher sum associated to* Γ, denoted $\bar{R}_\Gamma^{(-1)}(z)$, and defined by setting

$$\bar{R}_\Gamma^{(-1)}(z) = \mathrm{e}(-\bar{z}) + \lim_{K\to\infty}\sum_{\gamma\in(B(\mathbb{Z})\backslash\Gamma)^\times_{\leq K}}\mathrm{e}(-\gamma\cdot\bar{z}) - \mathrm{e}(-\gamma\cdot\infty) \tag{1.3.1}$$

for $z \in \mathbb{H}$. The conjugate Rademacher sum $\bar{R}_\Gamma^{(-1)}(z)$ defines an anti-holomorphic function on $\mathbb{H}$. At first glance it appears that we should recover the classical Rademacher sum $R_\Gamma^{(-1)}(z)$ after substituting $\bar{z}$ for z in $\bar{R}_\Gamma^{(-1)}(z)$, since the expressions defining $R_\Gamma^{(-1)}(z)$ and $\bar{R}_\Gamma^{(-1)}(\bar{z})$ would appear to coincide, but in fact the delicate limit defining the Rademacher sums behaves very differently depending on which half plane the variable z lies in, and the difference $R_\Gamma^{(-1)}(z) - \bar{R}_\Gamma^{(-1)}(\bar{z})$ can be rather far from vanishing. As an illustration of this, we show in §3.3 that the Fourier expansion of the conjugate Rademacher sum $\bar{R}_\Gamma^{(-1)}(z)$ in terms of the anti-holomorphic exponential $\mathrm{e}(\bar{z})$ recovers the values $\mathrm{c}_\Gamma(-1,n)$ for $n < 0$ as Fourier coefficients. More precisely, we establish the following counterpart to (1.2.10) (cf. Theorem 3.3.3).

$$\bar{R}_\Gamma^{(-1)}(z) = \mathrm{e}(-\bar{z}) - \frac{1}{2}\mathrm{c}_\Gamma(-1,0) - \sum_{n<0}\mathrm{c}_\Gamma(-1,n)\,\mathrm{e}(n\bar{z}) \tag{1.3.2}$$

We demonstrate in §3.4 that the holomorphic function $\bar{R}_\Gamma^{(-1)}(\bar{z})$ is also an abelian integral (cf. Theorem 3.4.4). More than this, the difference

$$R_\Gamma^{(-1)}(z) - \bar{R}_\Gamma^{(-1)}(z) \tag{1.3.3}$$

is a Γ-invariant harmonic function on $\mathbb{H}$ (cf. Theorem 3.4.6).

The Γ-invariance of the function (1.3.3) entails a remarkable formula for $\bar{R}_\Gamma^{(-1)}(z)$ in the case that Γ has genus zero. For since the holomorphic Rademacher sum $R_\Gamma^{(-1)}(z)$ is Γ-invariant in the genus zero case, the anti-holomorphic function $\bar{R}_\Gamma^{(-1)}(z)$ must also be Γ-invariant. Given our knowledge (1.3.2) of the Fourier expansion of $\bar{R}_\Gamma^{(-1)}(z)$ it follows that $\bar{R}_\Gamma^{(-1)}(z)$ must be identically constant, and furthermore, this constant must be $-\mathrm{c}_\Gamma(-1,0)/2$. In the case that $\Gamma = PSL_2(\mathbb{Z})$ we have $\mathrm{c}_\Gamma(-1,0) = 24$, and we thus obtain the remarkable identity

$$\mathrm{e}(-\bar{z}) + \lim_{K\to\infty} \sum_{\substack{0<c<K \\ -K^2<d<K^2 \\ (c,d)=1}} \mathrm{e}\left(-\frac{a\bar{z}+b}{c\bar{z}+d}\right) - \mathrm{e}\left(-\frac{a}{c}\right) = -12, \tag{1.3.4}$$

where in each summand a and b are integers chosen so that $ad - bc = 1$. This implies that the function $R_\Gamma^{(-1)}(z) - \bar{R}_\Gamma^{(-1)}(z)$ is a (non-normalized) hauptmodul for Γ when Γ has genus zero, and in particular,

$$\mathrm{c}_\Gamma(-1,n) = \delta_{-1,n} \tag{1.3.5}$$

for $n < 0$. We recover the normalized hauptmodul for Γ (cf. §2.5) by considering the sum $R_\Gamma^{(-1)}(z) + \bar{R}_\Gamma^{(-1)}(z)$.

1.4 Solid tori

It was observed in [CN79] that each group attached to the Monster via monstrous moonshine may be described as a group of $n\|h$-*type*, meaning a discrete group $\Gamma < PSL_2(\mathbb{R})$ of the form $\Gamma = \Gamma_0(n\|h) + S$ for some $n, h \in \mathbb{Z}^+$ with $h|(n, 24)$, and S a subgroup of the group of exact divisors of n/h. For such n, h, and S, the group $\Gamma_0(n\|h)+S$ contains and normalizes $\Gamma_0(nh)$, and has the property that $(\Gamma_0(n\|h) + S)_\infty = B(\mathbb{Z})$. (We recall the precise definition of $\Gamma_0(n\|h) + S$ in §2.5.)

We shed new light on the groups of $n\|h$-type by demonstrating in §6.2 that they solve a natural family of moduli problems for elliptic curves equipped with certain kinds of extra structure. As we will explain presently, a common feature of all the members of this family is a choice of filling of the elliptic curve in question; that is, a solid torus whose boundary is the given elliptic curve. Hence, we interpret the $n\|h$-type groups as defining moduli of (decorated)

solid tori.

The fact that each group Γ of $n\|h$-type satisfies $\Gamma_\infty = B(\mathbb{Z})$ suggests the importance of the quotient $B(\mathbb{Z})\backslash\mathbb{H}$. Regarding $\mathbb{H}$ as a moduli space for triples (E, γ, γ') where E is an elliptic curve and (γ, γ') is an oriented basis for the first homology group $H_1(E, \mathbb{Z})$ of E, we see that $B(\mathbb{Z})\backslash\mathbb{H}$ parameterizes pairs (E, γ) where E is an elliptic curve and γ is a primitive element of $H_1(E, \mathbb{Z})$. To each such pair (E, γ) is naturally associated an infinite volume hyperbolic 3-manifold with boundary an elliptic curve, for the cycle γ specifies a way to fill in the surface defined by E, thus yielding a solid torus. According to a theorem of Sullivan (cf. [Sul81], [McM92]) every complete, smooth, infinite volume hyperbolic 3-manifold with boundary an elliptic curve arises in this fashion, so we consider pairs (E, C) where E is an elliptic curve over $\mathbb{C}$ and C is an oriented subgroup of E isomorphic to S^1. We call such a pair a *solid torus*; according to our discussion the quotient space $B(\mathbb{Z})\backslash\mathbb{H}$ naturally parameterizes solid tori.

For (E, C) a solid torus and $n \in \mathbb{Z}^+$ we write $C[n]$ for the group of n-division points of C, and we define an *n-compatible isogeny of solid tori* $(E', C') \to (E, C)$ to be an isogeny $E' \to E$ of elliptic curves that maps $C'[n]$ to a subgroup of $C[n]$, and we define a *fully compatible isogeny of solid tori* to be an isogeny $E' \to E$ of elliptic curves that restricts to an orientation preserving map $C' \to C$ on the underlying primitive cycles. (Note that the n in an n-compatible isogeny of solid tori does not restrict the degree of the underlying isogeny of elliptic curves.) For (E, C) a solid torus, the canonical map $E \to E/C[n]$ may be viewed as defining a fully compatible isogeny of solid tori $(E, C) \to (E, C)/C[n]$, where $(E, C)/C[n]$ is a shorthand for the solid torus whose underlying elliptic curve is $E/C[n]$, and whose primitive cycle is the image of C under the natural map $E \to E/C[n]$. It may happen that an n-compatible isogeny induces invertible n/e-compatible isogenies (i.e. *n/e-compatible isomorphisms*) $(E', C')/C'[e] \to (E, C)$ and $(E', C') \to (E, C)/C[e]$ for some divisor e of n. In this case we say that (E', C') and (E, C) are *$n + e$-related*. By considering the existence or otherwise of such morphisms of solid tori we arrive at moduli interpretations for all the $n\|h$-type groups. For example, given a subgroup S of the group of exact divisors of n, say solid tori (E, C) and (E', C') are *$n + S$-related* if (E, C) and (E', C') are $n + e$-related for some $e \in S$. Then the $n + S$-relation is an equivalence relation on solid tori, and the quotient $(\Gamma_0(n) + S)\backslash\mathbb{H}$ is in natural correspondence with $n + S$-equivalence classes of solid tori.

In this way we demonstrate that the groups of $n\|h$-type may be explained as being the groups that solve a natural family of moduli problems for decorated solid tori with conformal structure on the boundary.

1.5 Moonshine via Rademacher sums

The fact that the function (1.2.5) recovers the normalized hauptmodul of the group Γ when Γ has genus zero is a strong indication that the Rademacher sum may be useful for the purpose of

understanding the functions of monstrous moonshine. A group theoretic characterization of the functions arising as McKay–Thompson series was given in [CMS04]. There the authors showed that a holomorphic function on $\mathbb{H}$ coincides with $T_g(z)$ for some $g \in \mathbb{M}$ if and only if it is the normalized hauptmodul for a group Γ satisfying each of the following properties.

1. The Riemann surface $\Gamma\backslash\mathbb{H}$ has genus zero.
2. The group Γ is of $n\|h$-type.
3. The group Γ is an extension of $\Gamma_0(nh)$ by a group of exponent 2.
4. Each cusp of Γ can be mapped to the infinite cusp by an element $\sigma \in PSL_2(\mathbb{R})$ with the property that $(\sigma\Gamma\sigma^{-1})_\infty = B(\mathbb{Z})$ and the intersection $\Gamma \cap \sigma\Gamma\sigma^{-1}$ contains $\Gamma_0(nh)$.

As we saw in §1.2 the variance of the Rademacher sum $R_\Gamma^{(-1)}(z)$ with respect to the action of Γ detects whether or not Γ has genus zero, and we have discussed in §1.4 the significance of the groups appearing in condition 2. We show in §§6.3,6.4 that the conditions 3 and 4 also admit natural reformulations in terms of the Rademacher sums.

Our reformulations come into view when we generalize the construction (1.2.2) so as to associate a Rademacher sum $R_{\Gamma,\mathsf{p}|\mathsf{q}}^{(-1)}(z)$ to each triple $(\Gamma, \mathsf{p}, \mathsf{q})$ where Γ is a group commensurable with $PSL_2(\mathbb{Z})$, and p and q are cusps of Γ. (This construction is presented in §3.1.) We write $R_{\Gamma,\mathsf{p}}^{(-1)}(z)$ as a shorthand for $R_{\Gamma,\mathsf{p}|\mathsf{q}}^{(-1)}(z)$ when q is the *infinite cusp* $\Gamma \cdot \infty$. It develops in §3.4 that $R_{\Gamma,\mathsf{p}}^{(-1)}(z)$ is an abelian integral for Γ with a simple pole at p, and no other singularities, and the expansion of $R_{\Gamma,\mathsf{p}|\mathsf{q}}^{(-1)}(z)$ at the infinite cusp coincides with that of $R_{\Gamma,\mathsf{p}}^{(-1)}(z)$ at q, up to (addition by) a certain constant function (cf. Theorem 3.4.4). Considering all the functions $R_{\Gamma,\mathsf{p}|\mathsf{q}}^{(-1)}(z)$ for varying cusps p and q we are able to access more subtle properties of the curve $\mathsf{X}_\Gamma = \Gamma\backslash\mathbb{H} \cup \Gamma\backslash\hat{\mathbb{Q}}$ (cf. §2.5) beyond its genus. We find in §6.3 that condition 3 translates into a certain condition of symmetry (cf. Proposition 6.3.2) in the Rademacher sums associated to the Hecke congruence group $\Gamma_0(nh)$, where n and h are as in condition 2, and we find in §6.4 that condition 4 translates into the condition (cf. Proposition 6.4.3) that the function $R_{\Gamma,\mathsf{p}|\mathsf{q}}^{(-1)}$ be invariant for this subgroup $\Gamma_0(nh)$.

Our moduli interpretation of the $n\|h$-type groups, together with the invariance and symmetry conditions on Rademacher sums just described, facilitate a reformulation of the conditions of [CMS04]. We thus arrive at a new characterization (cf. Theorem 6.5.1) of the functions of monstrous moonshine in terms of Rademacher sums, and moduli of solid tori.

1.6 Modified Rademacher sums

In order to understand better the behavior of the Rademacher sums $R_\Gamma^{(-1)}(z)$, and, in particular, the subtraction of the constant $\mathrm{e}(-\gamma \cdot \infty)$ in (1.2.2), we consider (also in §3) a generalization of the Rademacher sum adapted to the problem of constructing modular forms of arbitrary even

integer weight. Our approach to Rademacher sums of negative weight is inspired by the work of Niebur in [Nie74].

For $\kappa, m \in \mathbb{Z}$ we define the *Rademacher sum of weight* 2κ *and order* m *associated to* Γ by setting

$$R_{\Gamma}^{\kappa(m)}(z) = \mathrm{e}(mz) + \lim_{K\to\infty} \sum_{\gamma\in(B(\mathbb{Z})\backslash\Gamma)^{\times}_{\leq K}} \mathrm{e}(m\gamma\cdot z)\,\mathrm{r}^{\kappa}(m,\gamma,z)(cz+d)^{-2\kappa}, \tag{1.6.1}$$

where $c, d \in \mathbb{R}$ are chosen (for each summand) so that γ is the image of a matrix $\left(\begin{smallmatrix} a & b \\ c & d\end{smallmatrix}\right) \in SL_2(\mathbb{R})$ in $PSL_2(\mathbb{R})$, and $\mathrm{r}^{\kappa}(m,\gamma,z)$ is the *Rademacher regularization factor of weight* 2κ, satisfying

$$\mathrm{r}^{\kappa}(m,\gamma,z) = 1 - \mathrm{e}(m\gamma\cdot\infty - m\gamma\cdot z)\,\mathrm{e}(m\gamma\cdot z - m\gamma\cdot\infty)_{<1-2\kappa}, \tag{1.6.2}$$

where $\mathrm{e}(z)_{<K}$ denotes the *partial exponential function* $\mathrm{e}(z)_{<K} = \sum_{0\leq k<K}(2\pi\mathbf{i}z)^k/k!$ (cf. §3.1). Observe that we recover the Rademacher sum $R_{\Gamma}^{(-1)}(z)$ of (1.2.2) when $\kappa = 0$. Also, we have $\mathrm{r}^{\kappa}(m,\gamma,z) = 1$ when κ is positive, so that the Rademacher sum $R_{\Gamma}^{\kappa(m)}(z)$ is a holomorphic Poincaré series (cf. §3.1) when $\kappa > 0$ (which is absolutely and locally uniformly convergent for $\kappa > 1$).

In §3.1 we furnish generalizations of the functions $\mathrm{Kl}_{\gamma}(m,n)$ and $\mathrm{Bl}_{\gamma}(m,n)$ for arbitrary $\kappa \in \mathbb{Z}$, obtaining a generalization of the formula (1.2.6), and leading to the following analogue (cf. Theorem 3.3.2) of (1.2.10).

$$R_{\Gamma}^{\kappa(m)}(z) = \mathrm{e}(mz) + \frac{1}{2}\mathrm{c}_{\Gamma}^{\kappa}(m,0) + \sum_{n>0}\mathrm{c}_{\Gamma}^{\kappa}(m,n)\,\mathrm{e}(nz) \tag{1.6.3}$$

The *conjugate Rademacher sum of weight* 2κ *and order* m *associated to* Γ is defined in direct analogy with the definition (1.3.1) of the conjugate Rademacher sum $\bar{R}_{\Gamma}^{(-1)}(z)$, which we presently recognize as the conjugate Rademacher sum of weight 0 and order 1 associated to Γ.

$$\bar{R}_{\Gamma}^{\kappa(m)}(z) = \mathrm{e}(m\bar{z}) + \lim_{K\to\infty} \sum_{\gamma\in(B(\mathbb{Z})\backslash\Gamma)^{\times}_{\leq K}} \mathrm{e}(m\gamma\cdot\bar{z})\,\mathrm{r}^{\kappa}(m,\gamma,\bar{z})(c\bar{z}+d)^{-2\kappa}. \tag{1.6.4}$$

We have the following analogue of (1.3.2), and counterpart to (1.6.3), which leads naturally to analogues of (1.3.3) for arbitrary $\kappa \in \mathbb{Z}$ (cf. Theorem 3.3.3).

$$\bar{R}_{\Gamma}^{\kappa(m)}(z) = \mathrm{e}(m\bar{z}) - \frac{1}{2}\mathrm{c}_{\Gamma}^{\kappa}(m,0) - \sum_{n<0}\mathrm{c}_{\Gamma}^{\kappa}(m,n)\,\mathrm{e}(n\bar{z}) \tag{1.6.5}$$

The variance of the functions $R_{\Gamma}^{\kappa(m)}(z)$ with respect to the natural (weight 2κ) action of the group Γ was described by Niebur in [Nie74] for the case that Γ has a single cusp. We verify

in §4.5 that for Γ an arbitrary group commensurable with the modular group $PSL_2(\mathbb{Z})$, the function $R_\Gamma^{\kappa(m)}(z) + \mathrm{c}_\Gamma^\kappa(m,0)/2$ defines an *automorphic integral of weight* 2κ *for* Γ (cf. §2.7) when $m < 0$, meaning that $R_\Gamma^{\kappa(m)}(z)$ is holomorphic on $\mathbb{H}$, possibly having poles at cusps of Γ, and we have

$$\left(R_\Gamma^{\kappa(m)}(\gamma\cdot z) + \frac{1}{2}\mathrm{c}_\Gamma^\kappa(m,0)\right)(cz+d)^{-2\kappa} = R_\Gamma^{\kappa(m)}(z) + \frac{1}{2}\mathrm{c}_\Gamma^\kappa(m,0) + \omega(\gamma)(z) \tag{1.6.6}$$

for some function $\omega : \Gamma \to \mathbb{C}[z]$, where the polynomial $\omega(\gamma)(z)$ has degree at most -2κ in z (cf. Theorem 4.5.1, Proposition 4.4.10). An identification of the function ω may be given in terms of a certain canonically defined map $I_\kappa(\Gamma) \to S_{1-\kappa}(\Gamma)$, where $I_\kappa(\Gamma)$ denotes the space of automorphic integrals of weight 2κ for Γ (cf. §2.7), and $S_\kappa(\Gamma)$ denotes the space of cusp forms of weight 2κ for Γ. We show (cf. Theorem 4.5.4) that the sequence

$$0 \to M_\kappa(\Gamma) \to I_\kappa(\Gamma) \to S_{1-\kappa}(\Gamma) \to 0 \tag{1.6.7}$$

is exact, where $M_\kappa(\Gamma)$ denotes the space of modular forms of weight 2κ for Γ possibly having poles at cusps (cf. §2.7), and the map $M_\kappa(\Gamma) \to I_\kappa(\Gamma)$ is the natural inclusion.

Apart from the delicate convergence of the limit defining the Rademacher sum $R_\Gamma^{\kappa(m)}(z)$, the most curious feature of the functions $R_\Gamma^{\kappa(m)}(z)$ is that they transform naturally with respect to the group Γ only after the addition of the constant $\mathrm{c}_\Gamma^\kappa(m,0)/2$, which is typically non-zero. We overcome this feature of the classical Rademacher sum by introducing in §4 the *continued Rademacher sum of weight* 2κ *and order* m *associated to* Γ, defined for $\kappa \leq 0$ and $m < 0$ by setting

$$T_\Gamma^{\kappa(m)}(z,s) = \mathrm{e}(mz) + \sum_{\gamma\in(B(\mathbb{Z})\backslash\Gamma)^\times} \mathrm{e}(m\gamma\cdot z)\,\mathrm{t}^\kappa(m,\gamma,z,s)(cz+d)^{-2\kappa} \tag{1.6.8}$$

where $\mathrm{t}^\kappa(m,\gamma,z,s)$ is a generalization of the Rademacher regularization factor $\mathrm{r}^\kappa(m,\gamma,z)$ of (1.6.2) satisfying $\mathrm{t}^\kappa(m,\gamma,z,1) = \mathrm{r}^\kappa(m,\gamma,z)$.

The right hand side of (1.6.8) converges absolutely and locally uniformly in z and s for $z \in \mathbb{H}$ and $\Re(s) > 1$. We define $T_\Gamma^{\kappa(m)}(z)$ by taking the limit as s tends to 1 in $T_\Gamma^{\kappa(m)}(z,s)$. We find that

$$T_\Gamma^{\kappa(m)}(z) = \lim_{s\to 1^+} T_\Gamma^{\kappa(m)}(z,s) = R_\Gamma^{\kappa(m)}(z) - \frac{1}{2}\mathrm{c}_\Gamma^\kappa(m,0) \tag{1.6.9}$$

(cf. Proposition 4.4.9), so that the Fourier expansion of $T_\Gamma^{\kappa(m)}(z)$ has vanishing constant term. Then, in the case that Γ has genus zero, the function $T_\Gamma^{(-1)}(z) = T_\Gamma^{0(1)}(z)$ is precisely the normalized hauptmodul of Γ. We call $T_\Gamma^{\kappa(m)}(z)$ the *normalized Rademacher sum of weight* 2κ *and order* m *associated to* Γ. We may regard the disappearance of the constant term in (1.6.9)

as a consequence of the noncommutativity of the limits $s \to 1$ and $K \to \infty$ when applied to (the expression obtained by replacing $(B(\mathbb{Z})\backslash\Gamma)^{\times}$ with $(B(\mathbb{Z})\backslash\Gamma)^{\times}_{\leq K}$ in) the right hand side of (1.6.8).

To recover functions with natural modular properties at non-positive weights we modify the continued Rademacher sum by subtracting a renormalized value at $z = 0$.

$$Q_{\Gamma}^{\kappa(m)}(z,s) = T_{\Gamma}^{\kappa(m)}(z,s) - T_{0,\Gamma}^{\kappa(m)}(s) \tag{1.6.10}$$

The function $T_{0,\Gamma}^{\kappa(m)}(s)$ is defined by an expression analogous to (1.6.8) which also converges absolutely and locally uniformly in s for $\Re(s) > 1$, and would vanish at $s = 1$ if it were not for the noncommutativity of limits as $s \to 1$ and $K \to \infty$. We define the *modified Rademacher sum of weight* 2κ *and order* m *associated to* Γ, denoted $Q_{\Gamma}^{\kappa(m)}(z)$, by taking the limit as s tends to 1 in $Q_{\Gamma}^{\kappa(m)}(z,s)$, and as a counterpart to (1.6.9) we have

$$Q_{\Gamma}^{\kappa(m)}(z) = \lim_{s\to 1^+} Q_{\Gamma}^{\kappa(m)}(z,s) = R_{\Gamma}^{\kappa(m)}(z) + \frac{1}{2}\mathrm{c}_{\Gamma}^{\kappa}(m,0) \tag{1.6.11}$$

(cf. Proposition 4.4.10), so that, in light of (1.6.6), the modified Rademacher sum $Q_{\Gamma}^{\kappa(m)}(z)$ is an automorphic integral of weight 2κ for Γ.

The definition of the modified Rademacher sum $Q_{\Gamma}^{\kappa(m)}(z)$ is inspired by Hurwitz's relation for the Hurwitz zeta function, and our continuation procedure may be regarded as identifying the factor $1/2$, appearing in the constant term of the Fourier expansion (1.6.3) of the classical Rademacher sum of non-positive weight, with -1 times the value of the Riemann zeta function $\zeta(s)$ at $s = 0$ (cf. §4.3).

In order to obtain and analyze spanning sets for spaces of automorphic integrals we attach modified Rademacher sums $Q_{\Gamma,\mathsf{p}|\mathsf{q}}^{\kappa(m)}(z)$ to each triple $(\Gamma,\mathsf{p},\mathsf{q})$ where Γ is a group commensurable with the modular group $PSL_2(\mathbb{Z})$ and p and q are cusps of Γ (cf. §4.1). The function $Q_{\Gamma}^{\kappa(m)}(z)$ is then recovered upon taking both p and q to be the *infinite cusp* $\Gamma\cdot\infty$. If we write $Q_{\Gamma,\mathsf{p}}^{\kappa(m)}(z)$ as a shorthand for $Q_{\Gamma,\mathsf{p}|\mathsf{q}}^{\kappa(m)}(z)$ when $\mathsf{q} = \Gamma\cdot\infty$, then, as is shown in §4.5, the modified Rademacher sums $Q_{\Gamma,\mathsf{p}}^{\kappa(m)}(z)$, for varying cusps $\mathsf{p} \in \Gamma\backslash\hat{\mathbb{Q}}$ and positive integers $m \in \mathbb{Z}^+$, constitute a basis for the space $I_{\kappa}(\Gamma)$ when $\kappa < 0$ (cf. Theorem 4.5.3). In the case that $\kappa = 0$ the modified Rademacher sums $Q_{\Gamma,\mathsf{p}}^{\kappa(m)}(z)$ span a subspace of $I_0(\Gamma)$ of codimension 1, and a full basis is obtained by including a constant function. The function $Q_{\Gamma,\mathsf{p}}^{\kappa(m)}(z)$ defines an automorphic integral with a single pole at the cusp p, and the function $Q_{\Gamma,\mathsf{p}|\mathsf{q}}^{\kappa(m)}(z)$ encodes the Fourier expansion of $Q_{\Gamma,\mathsf{p}}^{\kappa(m)}(z)$ at the cusp q, and no correction of constant terms is necessary for the validity of these statements.

1.7 Hecke operators

The original result of Rademacher [Rad39] identifies the first Rademacher sum $R_\Gamma^{(-1)}(z)$, for $\Gamma = PSL_2(\mathbb{Z})$, with the function $J(z)+12$, where $J(z)$ is the normalized hauptmodul for the modular group. In light of this it is natural to try to identify more general Rademacher sums in a similar fashion. In this section we will take $\kappa = 0$, but we will allow m to be an arbitrary non-zero integer. The identity

$$R_\Gamma^{(m)}(z) = \overline{\bar{R}_\Gamma^{(-m)}(z)} \tag{1.7.1}$$

follows immediately from the definitions of the classical and conjugate Rademacher sums, demonstrating that the conjugate Rademacher sums of negative order may be recovered from the classical Rademacher sums with positive order. On the other hand, one can show, as we did for the special case that $m=-1$, that $\bar{R}_\Gamma^{(m)}(z)$ is the constant function with constant value $-\mathrm{c}_\Gamma(m,0)/2$ in case $m<0$ and Γ has genus zero. When $\Gamma = PSL_2(\mathbb{Z})$ we have $\mathrm{c}_\Gamma(m,0) = 24\sigma(-m,1)$ for $m<0$, where $\sigma(n,1)$ is the sum of the divisors of n, so we have the following generalization of (1.3.4) for $m<0$.

$$\mathrm{e}(m\bar{z}) + \lim_{K\to\infty} \sum_{\substack{0<c<K \\ -K^2<d<K^2 \\ (c,d)=1}} \mathrm{e}\left(m\frac{a\bar{z}+b}{c\bar{z}+d}\right) - \mathrm{e}\left(m\frac{a}{c}\right) = -12\sigma(-m,1) \tag{1.7.2}$$

To obtain an expression for the classical Rademacher sums $R_\Gamma^{(m)}(z)$ with negative m we can analyze again its Fourier coefficients $\mathrm{c}_\Gamma(m,n)$. It follows from our results in §6.1 that

$$\mathrm{c}_\Gamma(m,n) = \delta_{m,n} \tag{1.7.3}$$

for $m,n<0$ in case Γ has genus zero. The resulting function $T_\Gamma^{(m)}(z) = R_\Gamma^{(m)}(z) - \mathrm{c}_\Gamma(m,0)/2$, satisfying the identity

$$T_\Gamma^{(m)}(z) + \frac{1}{2}\mathrm{c}_\Gamma(m,0) = \mathrm{e}(mz) + \lim_{K\to\infty} \sum_{\gamma\in(B(\mathbb{Z})\backslash\Gamma)^\times_{\leq K}} \mathrm{e}(m\gamma\cdot z) - \mathrm{e}(m\gamma\cdot\infty), \tag{1.7.4}$$

is therefore holomorphic on the upper half plane, invariant for the action of Γ, and of the form $q^{-|m|}+o(1)$ for $q=\mathrm{e}(z)$. Consequently each normalized Rademacher sum $T_\Gamma^{(-m)}(z)$, for $m>0$, is expressible as a degree m polynomial in $T_\Gamma^{(-1)}(z) = T_\Gamma(z)$. This polynomial is none other than the so-called *m-th Faber polynomial for* $T_\Gamma(z)$.

Proceeding from another direction, we can obtain the functions $T_\Gamma^{(-m)}(z)$ from the normalized hauptmodul $T_\Gamma(z)$ by applying Hecke operators. In §5.5 we study, in particular, the case that Γ is the modular group $PSL_2(\mathbb{Z})$. Recall that for $n\in\mathbb{Z}^+$ the action of the Hecke operator $\hat{T}(n)$

on a modular function $f(z)$ may be given by setting

$$(\hat{T}(n)f)(z) = \frac{1}{n}\sum_{\substack{ad=n\\0\leq b<d}} f\left(\frac{az+b}{d}\right). \tag{1.7.5}$$

From this description one may deduce that the difference $T_\Gamma^{(-m)}(z) - m(\hat{T}(m)T_\Gamma)(z)$ is holomorphic in $\mathbb{H}$ and vanishes as $z \to \mathbf{i}\infty$. We thus obtain the remarkable expression

$$T_\Gamma^{(-m)}(z) = m(\hat{T}(m)T_\Gamma)(z) \tag{1.7.6}$$

for the higher order Rademacher sums which generalizes the original result of Rademacher. Combining the presentation (1.2.1) of $T_\Gamma(z)$ as a Rademacher sum with the definition (1.7.5) of the Hecke operator $\hat{T}(m)$ we obtain the expression

$$m(\hat{T}(m)T_\Gamma)(z) + \frac{1}{2}\mathrm{c}_\Gamma(-m,0) = \mathrm{e}(-mz) + \lim_{K\to\infty}\sum_{\gamma\in(B(\mathbb{Z})\backslash M(m))^\times_{\leq K}} \mathrm{e}(-\gamma\cdot z) - \mathrm{e}(-\gamma\cdot\infty), \tag{1.7.7}$$

where $M(m)$ denotes the (image in $PGL_2^+(\mathbb{Q})$ of the) set of 2×2 matrices with integral entries and determinant m. Comparison of (1.7.4) with (1.7.7), in view of the identity (1.7.6), suggests an equality of some, if not all, exponential terms in both sums, and in fact there is an injective map

$$\begin{aligned} B(\mathbb{Z})\backslash M(1) &\hookrightarrow B(\mathbb{Z})\backslash M(m)\\ \gamma &\mapsto \tilde{\gamma}, \end{aligned} \tag{1.7.8}$$

since $M(1) = \Gamma$ when $\Gamma = PSL_2(\mathbb{Z})$, with the property that $\mathrm{e}(-m\gamma\cdot z) = \mathrm{e}(-\tilde{\gamma}\cdot z)$. In the case that m is prime, the remaining terms assemble into a fractional power Rademacher sum

$$\mathrm{e}\left(-\frac{1}{m}z\right) + \lim_{K\to\infty}\sum_{\gamma\in(B(m\mathbb{Z})\backslash\Gamma)^\times_{\leq K}} \mathrm{e}\left(-\frac{1}{m}\gamma\cdot z\right) - \mathrm{e}\left(-\frac{1}{m}\gamma\cdot\infty\right), \tag{1.7.9}$$

where $B(m\mathbb{Z})$ is the *m-fold translation group*, generated by $z\mapsto z+m$. Thus we obtain another injective correspondence

$$\begin{aligned} B(m\mathbb{Z})\backslash M(1) &\hookrightarrow B(\mathbb{Z})\backslash M(m)\\ \gamma &\mapsto \tilde{\gamma} \end{aligned} \tag{1.7.10}$$

with the property that $\mathrm{e}(-\gamma\cdot z/m) = \mathrm{e}(-\tilde{\gamma}\cdot z)$. The identity (1.7.7) implies the vanishing of

the fractional power Rademacher sum (1.7.9). This can also be proven directly, and we give a general vanishing result for fractional Rademacher sums in §5.4. More generally we obtain fractional power sums of the form

$$\mathrm{e}\left(-\frac{m}{l^2}z\right)+\lim_{K\to\infty}\sum_{\gamma\in(B(l\mathbb{Z})\backslash\Gamma)^{\times}_{\leq K}}\mathrm{e}\left(-\frac{m}{l^2}\gamma\cdot z\right)-\mathrm{e}\left(-\frac{m}{l^2}\gamma\cdot\infty\right), \tag{1.7.11}$$

for each exact divisor l of m, and each of these fractional power sums vanishes except for the sum corresponding to $l=1$. Thus our analysis covers generalizations of the Rademacher sums to fractional orders in addition to the higher integral orders. All the results discussed in this section admit generalizations to arbitrary groups commensurable with the modular group but we restrict ourselves to the modular group in this article.

1.8 Monstrous Lie algebras

The presentation (1.7.6) of the functions $T_\Gamma^{(-m)}(z)$ by means of Hecke operators (we restrict to the case $\Gamma=PSL_2(\mathbb{Z})$ in §1.7) immediately implies that their Fourier coefficients are positive integers. This fact points to the existence of further algebraic structures beyond the moonshine module vertex operator algebra $V^\natural$. In his proof of the moonshine conjectures [Bor92] Borcherds introduced the Monster Lie algebra $\mathfrak{m}$, which admits a presentation as a bi-graded generalized Kac–Moody algebra. The key to Borcherds' method is the denominator identity for $\mathfrak{m}$. This in turn can be interpreted as a BGG-type resolution of the trivial $\mathfrak{m}$-module. Let $\mathcal{V}$ denote the Verma module with highest weight 0 for $\mathfrak{m}$. Then $\mathcal{V}$ is the first term in this BGG-type resolution of the trivial module, and the bi-graded dimension $\operatorname{gdim}\mathcal{V}$ of $\mathcal{V}$ is obtained by computing the coefficients of p and q in the expression

$$\operatorname{gdim}\mathcal{V}=\exp\left(\sum_{m>0}(\hat{T}(m)T_\Gamma)(z)p^m\right) \tag{1.8.1}$$

where $q=\mathrm{e}(z)$. Applying the operator $F\mapsto p\partial_p\log F$ to (1.8.1) we obtain a generating function for the higher (absolute) order normalized Rademacher sums $T_\Gamma^{(-m)}(z)$, by (1.7.6). It is well-known (cf. [Mac95]) that this operator relates the complete homogeneous symmetric functions to the power-sum symmetric functions. Thus the bi-graded dimension of the Verma module for the Monster Lie algebra may be viewed as the *complete Rademacher sum* (of weight 0) associated to Γ.

In §7.2 we consider a family of generalized Kac–Moody algebras $\{\mathfrak{m}_g\}$, indexed by elements g in the Monster group $\mathbb{M}$. Just as the Monster Lie algebra may be constructed from the moonshine module $V^\natural$, the monstrous Lie algebras $\mathfrak{m}_g$, studied by Carnahan in [Car07], may be constructed from the g-twisted $V^\natural$-modules $V_g^\natural$, for $g\in\mathbb{M}$. The graded dimensions of the spaces

$V_g^\natural$ are given by functions $J_g(z) = J_{\Gamma_g}(z)$, related to the McKay–Thompson series $T_g(z)$ via the involution $z \mapsto -1/z$.

$$J_g(z) = T_g(-1/z) \tag{1.8.2}$$

One should note that Borcherds set the precedent here, introducing a family $\{\mathfrak{m}'_g\}$ of monstrous Lie superalgebras in [Bor92]. These algebras $\mathfrak{m}'_g$ are more directly related to the functions $T_g(z)$, rather than the $J_g(z)$, and for our purposes the algebras $\mathfrak{m}_g$ appear to be more convenient.

The monstrous Lie algebras $\mathfrak{m}_g$ are constructed in [Car07] using the semi-infinite cohomology version of the no-ghost theorem. We identify the bi-graded subspaces of the algebra $\mathfrak{m}_g$ in terms of the twisted modules $V_h^\natural$, with $h \in \langle g \rangle$. One can show that the $\mathfrak{m}_g$ are generalized Kac–Moody algebras, and one can also deduce formulas for the bi-graded dimensions of their Verma modules $\mathcal{V}_g$, generalizing (1.8.1).

$$\operatorname{gdim} \mathcal{V}_g = \exp\left(\sum_{m>0} (\hat{T}(m)T_g)(z)p^m\right) \tag{1.8.3}$$

Carnahan [Car07] also found a remarkable denominator identity which yields an alternative expression for the graded dimension of $\mathcal{V}_g$; namely,

$$\operatorname{gdim} \mathcal{V}_g = \frac{1}{p(T_g(w) - J_g(z))} \tag{1.8.4}$$

where $p = \mathrm{e}(w)$. This identity (1.8.4) demonstrates that the bi-graded dimension can be viewed as a meromorphic function on $\mathbb{H} \times \mathbb{H}$ with poles at $w \in \Gamma_g \cdot (-1/z)$.

We also show in §7.2 that the bi-graded dimension of $\mathcal{V}_g$ is bounded below, coefficient-wise, by the generating function

$$\sum_{m>0} J_g^{(-m)}(z)p^m \tag{1.8.5}$$

where $J_g^{(-m)}(z) = T_g^{(-m)}(-1/z)$ when g is of Fricke type (cf. Proposition 7.2.4). This suggests that the Verma modules $\mathcal{V}_g$ may contain naturally defined subspaces whose bi-graded dimensions coincide with the expressions (1.8.5).

1.9 Chiral gravity

Our results on Rademacher sums, at first glance, add as much to the mystery of monstrous moonshine as they reveal. We have formulated a characterization of the groups of monstrous moonshine in terms of Rademacher sums and solid tori. Having done so we face the new question of where the Rademacher sums themselves appear in relation to the moonshine module vertex

operator algebra $V^\natural$? It appears now that the relation to physics should again be useful, as it was in the construction of $V^\natural$ [FLM88], and the proof of the moonshine conjecture [Bor92].

In §7.1 we explain how the Rademacher sums associated to elements of the Monster might be regarded as giving strong evidence for the existence of another construction of the moonshine module $V^\natural$, and its twisted sectors $V_g^\natural$, in which all the features of monstrous moonshine, including the genus zero property, become transparent. In fact the original Rademacher sum (1.2.1) does appear as a saddle point approximation to the partition function of the simplest chiral three dimensional quantum gravity. (See [Wit07], [MW07], [Man07], [MM07], [LSS08] and [MSS09] for the development of this idea in the physics literature.) Our new identity (1.3.4) for the conjugate Rademacher sum must reflect the chiral nature of this three dimensional quantum gravity.

We expect that our analytic continuation of the Rademacher sums, and the explanation of the appearance of the constant term, will also develop from a rigorous analysis of the saddle point approximation using a zeta function regularization of the chiral gravity partition function, as was done for various models of two dimensional CFT. Our analysis of the Rademacher sums corresponding to the discrete groups of monstrous moonshine, and of moduli spaces of decorated solid tori points to a description of all the spaces $V_g^\natural$, for $g \in \mathbb{M}$, via a family of g-twisted versions of the simplest chiral three dimensional quantum gravity. We expect that our reformulation of monstrous moonshine in terms of Rademacher sums and moduli of solid tori will eventually be understood from the properties of this remarkable chiral three dimensional quantum gravity.

In §7.3 we also elucidate the relationship between the higher order Rademacher sums, and the key object of Borcherds' proof of the moonshine conjecture; viz., the Monster Lie algebra, and its twisted counterparts. This leads us to the stringy quantization (cf. [DMVV97]) of $V^\natural$, and we generalize this to all twisted sectors $V_g^\natural$, for $g \in \mathbb{M}$. We prove in Theorem 7.3.1 that the resulting second quantized spaces have the same bi-graded dimensions as the Verma modules $\mathcal{V}_g$ (cf. (1.8.3)) and thus may be viewed as stringy realizations of the representations of the generalized Kac–Moody algebras $\mathfrak{m}_g$. It is an interesting problem to interpret the spaces $\mathcal{V}_g$, and their subspaces with graded dimension given by the higher order normalized Rademacher sums (1.8.5), in terms of the second quantized chiral three dimensional quantum gravities. In particular, the simple description of the singularities of the partition function given by (1.8.4) should have a natural interpretation in terms of quantum gravity.

Thus our present results on the Rademacher sums, and the new conjectures about their origin, clearly indicate that the real nature of moonshine might not be as remote anymore, and more than this, the full examination of its structure might give new insight into the fundamental problem of modern physics.

2 Conventions

We write $\mathbb{Z}^+$ for the set of positive integers, and the notations $\mathbb{Q}^+$ and $\mathbb{R}^+$ are to be interpreted similarly. We write $\mathbb{N}$ for the set of non-negative integers. For R a ring without zero divisors, we write $R^\times$ for the multiplicative monoid of non-zero elements in R.

2.1 Functions

For $z, s \in \mathbb{C}$ with $z \neq 0$, we write z^s as a shorthand for $\exp(s\log(z))$, where log denotes the principal branch of the logarithm, so that

$$-\pi < \Im(\log(z)) \leq \pi, \tag{2.1.1}$$

and we write $z^{(s)}$ as a shorthand for $z^s/\Gamma(s+1)$, where $\Gamma(s)$ denotes the Gamma function. We adopt the convention of setting

$$\mathrm{e}(z) = \exp(2\pi\mathbf{i}z) = \sum_{k\geq 0}(2\pi\mathbf{i}z)^{(k)} \tag{2.1.2}$$

for $z \in \mathbb{C}$. We write $\Phi(a,b,z)$ for the analytic function on $\mathbb{C}^3$ defined by setting

$$\Phi(a,b,z) = \frac{{}_1F_1(a;b;2\pi\mathbf{i}z)}{\Gamma(b)} = \sum_{k\geq 0}\frac{\Gamma(k+a)}{\Gamma(a)\Gamma(k+b)}(2\pi\mathbf{i}z)^{(k)} \tag{2.1.3}$$

where ${}_1F_1(a;b;x)$ denotes the confluent hypergeometric function (cf. [AS64, §13]). Kummer's transformations for the confluent hypergeometric function yield the following identity for Φ.

$$\Phi(a,b,z) = \mathrm{e}(z)\Phi(b-a,b,-z) \tag{2.1.4}$$

The exponential function $\mathrm{e}(z)$ is a solution to the differential equation

$$(z\partial_z + 1 - s)(\partial_z - 2\pi\mathbf{i})u(z) = 0, \tag{2.1.5}$$

which has a regular singular point at 0 and an irregular singularity at ∞. Another solution is furnished by the function $z \mapsto \Phi(1,1+s,z)(2\pi\mathbf{i}z)^s$, which we denote also by $\mathrm{e}(z,s)$.

$$\mathrm{e}(z,s) = \Phi(1,1+s,z)(2\pi\mathbf{i}z)^s = \sum_{k\geq 0}(2\pi\mathbf{i}z)^{(k+s)} \tag{2.1.6}$$

Observe that we have $\mathrm{e}(z,n) = \mathrm{e}(z) - \mathrm{e}(z)_{<n}$ for $n \in \mathbb{Z}$, where $\mathrm{e}(z)_{<K}$ denotes the *partial exponential* $\mathrm{e}(z)_{<K} = \sum_{0\leq k<K}(2\pi\mathbf{i}z)^{(k)}$. In particular, $\mathrm{e}(z,n) = \mathrm{e}(z)$ when $n \leq 0$.

2.2 Divisors

For $n \in \mathbb{Z}^+$ and $s \in \mathbb{C}$ we write $\sigma(n,s)$ for the *divisor function*

$$\sigma(n,s) = \sum_{d|n} d^s. \tag{2.2.1}$$

For $a, b \in \mathbb{Z}$ we write (a,b) for the greatest common divisor of a and b, and we generalize this notation by writing $((a,b)$ for the greatest positive integer c say, such that $c^2|a$ and $c|b$, and by writing (a^∞, b) for the largest divisor of b that divides some power of a.

$$((a,b) = \max\left\{c \in \mathbb{Z}^+ \mid c^2|a \text{ and } c|b\right\} \tag{2.2.2}$$

$$(a^\infty, b) = \lim_{n\to\infty} (a^n, b) \tag{2.2.3}$$

For $d, n \in \mathbb{Z}^+$ we write $d\|n$ in the case that $d|n$ and $(d, n/d) = 1$, and call such a d an *exact divisor of* n. We write $\mathrm{Ex}(n)$ for the set of exact divisors of n. Then $\mathrm{Ex}(n)$ becomes a group of exponent 2 when equipped with the product $(e,f) \mapsto ef/(e,f)^2$. Indeed, if we write $\Pi(n)$ for the set of primes that divide n then there is a naturally defined isomorphism $\mathrm{Ex}(n) \cong (\mathbb{Z}/2)^{\Pi(n)}$ which associates to an exact divisor $e \in \mathrm{Ex}(n)$ the indicator function $\chi_e : \Pi(n) \to \mathbb{Z}/2$, satisfying $\chi_e(p) = 1$ if $p|e$ and $\chi_e(p) = 0$ otherwise.

Let Π denote the set of all primes. Then the infinite product $(\mathbb{Z}/2)^\Pi$ comes equipped with a naturally defined projection $(\mathbb{Z}/2)^\Pi \to (\mathbb{Z}/2)^{\Pi(n)}$, and a naturally defined section for it $(\mathbb{Z}/2)^{\Pi(n)} \to (\mathbb{Z}/2)^\Pi$, for each $n \in \mathbb{Z}^+$. Consequently we have naturally defined morphisms of groups $(\mathbb{Z}/2)^\Pi \to \mathrm{Ex}(n)$ and $\mathrm{Ex}(n) \to (\mathbb{Z}/2)^\Pi$ for each n, and for any $m, n \in \mathbb{Z}^+$ we have the natural map $\mathrm{Ex}(m) \to \mathrm{Ex}(n)$ furnished by the composition $\mathrm{Ex}(m) \to (\mathbb{Z}/2)^\Pi \to \mathrm{Ex}(n)$. If m divides n then the map $\mathrm{Ex}(m) \to \mathrm{Ex}(n)$ is injective, the map $\mathrm{Ex}(n) \to \mathrm{Ex}(m)$ is surjective, and the composition $\mathrm{Ex}(m) \to \mathrm{Ex}(n) \to \mathrm{Ex}(m)$ is the identity on $\mathrm{Ex}(m)$.

2.3 Isometries

Let us write $G(\mathbb{R})$ for the simple real Lie group $PSL_2(\mathbb{R})$. The group $G(\mathbb{R})$ acts naturally, from the left, on the upper-half plane $\mathbb{H} = \{z \in \mathbb{C} \mid \Im(z) > 0\}$, and is just the group of orientation preserving isometries of $\mathbb{H}$ when we equip it with the *hyperbolic measure* $\mathrm{d}\mu(z) = \mathrm{d}x\mathrm{d}y/y^2$, for $z = x + \mathbf{i}y$.

It is convenient to enlist matrices in $SL_2(\mathbb{R})$ for the purpose of specifying elements of $G(\mathbb{R})$. We write $\left[\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right]$ for the image of a matrix $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in SL_2(\mathbb{R})$ in $G(\mathbb{R})$. Analogously, we write $A \mapsto [A]$ for the canonical map $SL_2(\mathbb{R}) \to G(\mathbb{R})$. The action of $G(\mathbb{R})$ on $\mathbb{H}$ is now described explicitly by

the formula

$$\gamma \cdot z = \frac{az+b}{cz+d} \tag{2.3.1}$$

for $\gamma = \left[\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right]$. The matrix $\theta = \left(\begin{smallmatrix} -1 & 0 \\ 0 & 1 \end{smallmatrix}\right) \in GL_2(\mathbb{R})$ induces an outer-automorphism of $G(\mathbb{R})$, which we call *conjugation on* $G(\mathbb{R})$, and which we denote $\gamma \mapsto \bar{\gamma}$. Explicitly, we have

$$\bar{\gamma} = \begin{bmatrix} -a & b \\ c & -d \end{bmatrix} \Longleftarrow \gamma = \begin{bmatrix} a & b \\ c & d \end{bmatrix}. \tag{2.3.2}$$

For each matrix $A \in GL_2^+(\mathbb{Q})$ there is a unique $\mu \in \mathbb{R}^+$ for which $A' = \mu A$ belongs to $SL_2(\mathbb{R})$. The assignment $GL_2^+(\mathbb{Q}) \to G(\mathbb{R})$ given by $A \mapsto [A']$ then factors through $PGL_2^+(\mathbb{Q})$, and the resulting map $PGL_2^+(\mathbb{Q}) \to G(\mathbb{R})$ is an embedding of groups. We write $G(\mathbb{Q})$ for the copy of $PGL_2^+(\mathbb{Q})$ in $G(\mathbb{R})$ obtained in this way. We write $G(\mathbb{Z})$ for the *modular group* $PSL_2(\mathbb{Z})$.

Extending the notation introduced above, we write $[A]$ and $\left[\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right]$ for the image in $G(\mathbb{Q}) < G(\mathbb{R})$ of a matrix $A = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$ in $GL_2^+(\mathbb{Q})$. Given $\mu \in \mathbb{Q}^+$ we write $[\mu]$ as a shorthand for $[A]$ when A is the diagonal matrix $\left(\begin{smallmatrix} \mu & 0 \\ 0 & 1 \end{smallmatrix}\right)$. Then $[\mu] \cdot z = \mu z$ for $z \in \mathbb{H}$.

$$[\mu] = \begin{bmatrix} \mu & 0 \\ 0 & 1 \end{bmatrix} \tag{2.3.3}$$

Let $B(\mathbb{R})$ denote the Borel subgroup of $G(\mathbb{R})$ consisting of images of upper-triangular matrices in $SL_2(\mathbb{R})$, so that $B(\mathbb{R})$ is just the subgroup of $G(\mathbb{R})$ that fixes the distinguished point ∞ on the *pointed real projective line* $\hat{\mathbb{R}} = \mathbb{R} \cup \{\infty\}$. Let $B(\mathbb{Q})$ and $B(\mathbb{Z})$ denote the intersections $B(\mathbb{R}) \cap G(\mathbb{Q})$ and $B(\mathbb{R}) \cap G(\mathbb{Z})$, respectively. Then $B(\mathbb{Z})$ is the group generated by the *modular translation* $T = \left[\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right]$. Set $\hat{\mathbb{Q}} = \mathbb{Q} \cup \{\infty\} \subset \hat{\mathbb{R}}$. Then $G(\mathbb{Q})$ acts naturally on $\hat{\mathbb{Q}}$, and $B(\mathbb{Q})$ is just the subgroup of $G(\mathbb{Q})$ that fixes the point ∞. For $\alpha \in \mathbb{Q}$ we write T^α for the element $\left[\begin{smallmatrix} 1 & \alpha \\ 0 & 1 \end{smallmatrix}\right] \in B(\mathbb{Q})$.

$$T^\alpha = \begin{bmatrix} 1 & \alpha \\ 0 & 1 \end{bmatrix} \tag{2.3.4}$$

Then the elements of $B(\mathbb{Q})$ of the form T^α for $\alpha \in \mathbb{Q}$ constitute the *unipotent subgroup of* $B(\mathbb{Q})$, which we denote $B_u(\mathbb{Q})$. Given $\alpha \in \mathbb{Q}^+$ we write $B(\alpha\mathbb{Z})$ for the subgroup of $B_u(\mathbb{Q})$ generated by T^α.

$$B(\alpha\mathbb{Z}) = \left\{ \begin{bmatrix} 1 & \alpha n \\ 0 & 1 \end{bmatrix} \mid n \in \mathbb{Z} \right\} \tag{2.3.5}$$

The elements of $B(\mathbb{Q})$ of the form $[\mu]$ for $\mu \in \mathbb{Q}^+$ constitute the *diagonal subgroup of* $B(\mathbb{Q})$, which we denote $B_d(\mathbb{Q})$. The group $B(\mathbb{Q})$ is naturally isomorphic to the semidirect product $B_u(\mathbb{Q}) \rtimes B_d(\mathbb{Q})$. In particular, $B_u(\mathbb{Q})$ is a normal subgroup of $B(\mathbb{Q})$ and for any $\chi \in B(\mathbb{Q})$ we have $\chi = T^\alpha[\mu]$ for some uniquely determined $\alpha \in \mathbb{Q}$ and $\mu \in \mathbb{Q}^+$.

An element $\gamma \in G(\mathbb{Q})$ will have many preimages in $GL_2^+(\mathbb{Q})$, but among these there will be exactly two that have integral entries with no common divisor. Either of these two matrices will be called a *preferred representative* for γ. We define functions $P\det : G(\mathbb{Q}) \to \mathbb{Z}^+$ and $c : G(\mathbb{Q}) \to \mathbb{N}$ and $d : G(\mathbb{Q}) \to \mathbb{N}$ by setting

$$P\det(\gamma) = ad - bc, \quad c(\gamma) = |c|, \quad d(\gamma) = |d|, \tag{2.3.6}$$

in case $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in GL_2^+(\mathbb{Q})$ is a preferred representative for γ. We call $P\det$ the *projective determinant*. The group $G(\mathbb{Z})$ is exactly the preimage of 1 with respect to the projective determinant. The projective determinant is not multiplicative, but we have $P\det(\gamma\sigma) = P\det(\sigma) = P\det(\sigma\gamma)$ for all $\sigma \in G(\mathbb{Q})$ in case $P\det(\gamma) = 1$, so for any $n \in \mathbb{Z}^+$ the preimage of n under $P\det$ admits commuting left and right actions of $G(\mathbb{Z})$.

Observe that $c(T\gamma) = c(\gamma T) = c(\gamma)$ and $d(T\gamma) = d(\gamma)$. Given $X \subset G(\mathbb{Q})$ we set $X_\infty = X \cap B(\mathbb{Q})$ and $X^\times = X - X_\infty$. Then $X_\infty = \{\chi \in X \mid c(\chi) = 0\}$. We define $X^{\times\times}$ to be the subset of $X^\times$ consisting of $\chi \in X$ such that both $c(\chi)$ and $d(\chi)$ are non-zero.

$$X_\infty = \{\chi \in X \mid c(\chi) = 0\}, \tag{2.3.7}$$

$$X^\times = \{\chi \in X \mid c(\chi) \neq 0\}, \quad X^{\times\times} = \{\chi \in X \mid c(\chi)d(\chi) \neq 0\} \tag{2.3.8}$$

For $\gamma \in G(\mathbb{Q})$ with $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$ a preferred representative, we have

$$\gamma \cdot z - \gamma \cdot z' = \frac{P\det(\gamma)}{(cz+d)(cz'+d)}(z - z') \tag{2.3.9}$$

for $z, z' \in \mathbb{H}$, so that the derivative of the function $z \mapsto \gamma \cdot z$, which we will denote $\mathrm{jac}(\gamma, z)$, is given by

$$\mathrm{jac}(\gamma, z) = \frac{P\det(\gamma)}{(cz+d)^2}. \tag{2.3.10}$$

Note that the assignment $\mathrm{jac} : G(\mathbb{Q}) \to \mathcal{O}(\mathbb{H})$, which associates the function $z \mapsto \mathrm{jac}(\gamma, z)$ to an element $\gamma \in G(\mathbb{Q})$, descends naturally to the coset space $B(\mathbb{Z})\backslash G(\mathbb{Q})$, and even further to $B_u(\mathbb{Q})\backslash G(\mathbb{Q})$.

From (2.3.9) and (2.3.10) we see that $|\gamma \cdot z - \gamma \cdot z'| = |z - z'|$ just in the case that z and z' belong to the set $\{z \in \mathbb{H} \mid |\mathrm{jac}(\gamma, z)| = 1\}$ which we call the *isometric locus of* γ. The isometric locus of γ is a (Euclidean) semicircle just when $c(\gamma) > 0$, in which case its center is $\gamma^{-1} \cdot \infty$,

and its radius is $\sqrt{\varrho(\gamma)}$, where

$$\varrho(\gamma) = \frac{P\det(\gamma)}{c(\gamma)^2} = \frac{ad-bc}{c^2} \tag{2.3.11}$$

in case $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$ is a preferred representative for γ. We call $\varrho(\gamma)$ the *scaling factor* associated to γ.

2.4 Cosets

Given a subset $X \subset G(\mathbb{Q})$, we will write $[\![X]\!]$ for the set consisting of double cosets of the form $B(\mathbb{Z})\chi B(\mathbb{Z})$, for $\chi \in X$. We will write $[\![\chi]\!]$ as a shorthand for the particular double coset $B(\mathbb{Z})\chi B(\mathbb{Z})$.

$$[\![\chi]\!] = B(\mathbb{Z})\chi B(\mathbb{Z}), \quad [\![X]\!] = \{[\![\chi]\!] \in B(\mathbb{Z})\backslash G(\mathbb{Q})/B(\mathbb{Z}) \mid \chi \in X\} \tag{2.4.1}$$

Similarly, we will write $[\![X$ for the set of right cosets of the form $B(\mathbb{Z})\chi$, for $\chi \in X$, and $\chi \mapsto [\![\chi$ will denote the natural map $X \to [\![X$, and the notations $X]\!]$ and $\chi \mapsto \chi]\!]$ will have the analogous meanings. Observe that $B(\mathbb{Z})$ is stable under the operations of inversion and conjugation (cf. (2.3.2)), and thus these operations descend naturally to the coset spaces $[\![G(\mathbb{Q})$ and $G(\mathbb{Q})]\!]$ and $[\![G(\mathbb{Q})]\!]$.

Recall from §2.3 that the assignment $\chi \mapsto \mathrm{jac}(\chi, z)$ descends to a well-defined association of holomorphic functions on $\mathbb{H}$ to right cosets of $B(\mathbb{Z})$ in $G(\mathbb{Q})$. Accordingly, we may safely set $\mathrm{jac}([\![\chi\,, z) = \mathrm{jac}(\chi, z)$, when given some $[\![\chi \in [\![G(\mathbb{Q})$.

$$\begin{aligned} \mathrm{jac}(\cdot\,, z) : [\![G(\mathbb{Q}) &\to \mathcal{O}(\mathbb{H}) \\ [\![\chi &\mapsto \mathrm{jac}([\![\chi\,, z) \end{aligned} \tag{2.4.2}$$

Given $U \subset [\![G(\mathbb{Q})$ or $S \subset [\![G(\mathbb{Q})]\!]$ we set $U_\infty = U \cap [\![B(\mathbb{Q})$ and $U^\times = U - U_\infty$, and similarly, $S_\infty = S \cap [\![B(\mathbb{Q})]\!]$ and $S^\times = S - S_\infty$. In particular, $[\![G(\mathbb{Q})]\!]^\times$ denotes the set of double cosets $[\![\chi]\!]$, for $\chi \in G(\mathbb{Q})$, for which $c(\chi) \neq 0$.

Observe that the isometric loci (cf. §2.3) associated to $\chi, \chi' \in G(\mathbb{Q})$ coincide when $[\![\chi = [\![\chi'$. Thus we may regard these loci as naturally attached to right cosets of $B(\mathbb{Z})$ in $G(\mathbb{Q})$, and the locus attached to $[\![\chi$ is a Euclidean semicircle just when $[\![\chi \in [\![G(\mathbb{Q})^\times$. Also, the loci associated to cosets $[\![\chi\,, [\![\chi' \in [\![G(\mathbb{Q})^\times$ have the same radii when $[\![\chi]\!] = [\![\chi']\!]$; that is to say, the function ϱ of (2.3.11) descends to a well-defined map $[\![G(\mathbb{Q})]\!]^\times \to \mathbb{Q}^+$. It is worth noting that the function $\chi \mapsto \varrho(\chi)$ actually satisfies an even stronger invariance condition.

Lemma 2.4.1. *The assignment $[\![\chi]\!] \mapsto \varrho\,[\![\chi]\!]$ descends naturally to a well-defined function on the double coset space $B_u(\mathbb{Q})\backslash G(\mathbb{Q})/B_u(\mathbb{Q})$. That is, we have $\varrho\,[\![T^\alpha\chi]\!] = \varrho\,[\![\chi T^\alpha]\!] = \varrho\,[\![\chi]\!]$ for any $\alpha \in \mathbb{Q}$ and $\chi \in G(\mathbb{Q})^\times$.*

The function ϱ is also invariant under inversion and conjugation, so that we have $\varrho\,[\![\chi]\!] = \varrho\,[\![\bar{\chi}]\!] = \varrho\,[\![\chi^{-1}]\!]$ for any $\chi \in G(\mathbb{Q})^\times$. For $\chi \in G(\mathbb{Q})^\times$ the identity (2.3.9) may be rewritten

$$\chi \cdot z - \chi \cdot z' = \frac{\varrho\,[\![\chi]\!]}{(z - \chi^{-1} \cdot \infty)(z' - \chi^{-1} \cdot \infty)}\ (z - z'), \tag{2.4.3}$$

and we may consider the limit as $z' \to \infty$, which yields a useful expression for $\chi \cdot z - \chi \cdot \infty$; viz.,

$$\chi \cdot z - \chi \cdot \infty = -\frac{\varrho\,[\![\chi]\!]}{z - \chi^{-1} \cdot \infty}. \tag{2.4.4}$$

Comparing with (2.3.10) and (2.4.4) we have

$$\mathrm{jac}([\![\chi\,, z) = \frac{\varrho\,[\![\chi]\!]}{(z - \chi^{-1}]\!] \cdot \infty)^2} \tag{2.4.5}$$

for $[\![\chi \in [\![G(\mathbb{Q})^{\times}$ (and $\mathrm{jac}([\![\chi\,, z) = 1$ otherwise).

The functions $c, d : G(\mathbb{Q}) \to \mathbb{N}$, of §2.3, descend to well-defined functions on the coset space $[\![G(\mathbb{Q}) = B(\mathbb{Z})\backslash G(\mathbb{Q})$. The function c descends further, to be well-defined on the double coset space $[\![G(\mathbb{Q})]\!] = B(\mathbb{Z})\backslash G(\mathbb{Q})/B(\mathbb{Z})$. For $K \in \mathbb{R}^+$ and U a subset of $[\![G(\mathbb{Q})$, define $U_{\leq K}$ to be the subset of U consisting of cosets $[\![\chi \in U$ for which $c(\chi) \leq K$ and $d(\chi) \leq K^2$. Analogously, for S a subset of $[\![G(\mathbb{Q})]\!]$, define $S_{\leq K}$ to be the subset of S consisting of double cosets $[\![\chi]\!] \in S$ for which $c(\chi) \leq K$.

$$U_{\leq K} = \left\{[\![\chi \in U \mid c(\chi) \leq K,\, d(\chi) \leq K^2\right\} \tag{2.4.6}$$

$$S_{\leq K} = \{[\![\chi]\!] \in S \mid c(\chi) \leq K\} \tag{2.4.7}$$

2.5 Groups

The group $G(\mathbb{Q})$ is the commensurator of $G(\mathbb{Z})$. Consequently, any group $\Gamma < G(\mathbb{R})$ that is commensurable with the modular group $G(\mathbb{Z})$ is automatically a subgroup of $G(\mathbb{Q})$, and is an example of a Fuchsian group of the first kind. For such a group Γ there is a natural way to equip the orbit space $\Gamma\backslash\mathbb{H}$ with the structure of a Riemann surface; we will denote this object by Y_Γ. Any group Γ commensurable with $G(\mathbb{Z})$ has parabolic elements, and any fixed point of any parabolic element of Γ lies on the (pointed) rational projective line $\hat{\mathbb{Q}} \subset \hat{\mathbb{R}}$. The orbit space $\mathsf{P}_\Gamma = \Gamma\backslash\hat{\mathbb{Q}}$ is the set of *cusps of* Γ. For Γ commensurable with $G(\mathbb{Z})$, the Riemann surface Y_Γ is not compact, but can be compactified in a natural way by adjoining a single point for each cusp of Γ. We will write X_Γ for the corresponding compact Riemann surface obtained by the adjunction of the cusps of Γ.

$$\mathsf{X}_\Gamma = \mathsf{Y}_\Gamma \cup \mathsf{P}_\Gamma = \Gamma\backslash\mathbb{H} \cup \Gamma\backslash\hat{\mathbb{Q}} \tag{2.5.1}$$

We will say that Γ is a *group of genus zero* in the case that X_Γ has genus zero as a Riemann surface. We set $\mathrm{genus}(\Gamma) = \mathrm{genus}(\mathsf{X}_\Gamma)$ for $\Gamma < G(\mathbb{R})$ commensurable with $G(\mathbb{Z})$.

We write $\hat{\mathbb{C}}$ for the *(pointed) complex projective line* $\mathbb{C} \cup \{\infty\}$, which we regard as a compact Riemann surface in the usual way. If Γ is a group for which the compact Riemann surface X_Γ has genus zero, then there is an isomorphism $\phi : \mathsf{X}_\Gamma \xrightarrow{\sim} \hat{\mathbb{C}}$ that witnesses this fact, and the set of such isomorphisms admits a simply transitive action by $G(\mathbb{C})$. Thus we may assume that ϕ maps the point of X_Γ corresponding to the *infinite cusp* $\Gamma \cdot \infty$ to the distinguished point $\infty \in \hat{\mathbb{C}}$. Such an isomorphism ϕ determines a Γ-invariant holomorphic function, f say, on $\mathbb{H}$, which admits an expression $f(z) = \sum_{n \geq -1} c(n/h)\, \mathrm{e}(nz/h)$ with $c(-1/h) \neq 0$ for some positive integer h called the *width of* Γ *at infinity*. After multiplying by $1/c(-1/h)$ we may assume that $c(-1/h) = 1$. A Γ-invariant function of the form $\mathrm{e}(-z/h) + \sum_{n \geq 0} c(n/h)\, \mathrm{e}(nz/h)$ which is holomorphic on $\mathbb{H}$ and induces an isomorphism $\mathsf{X}_\Gamma \xrightarrow{\sim} \hat{\mathbb{C}}$ will be called a *hauptmodul* for Γ. In case $c(0) = 0$ we say that f is a *normalized hauptmodul* for Γ.

The *Hecke congruence groups*, denoted $\Gamma_0(n)$ for $n \in \mathbb{Z}^+$, play a special rôle in our analysis.

$$\Gamma_0(n) = \left\{ \begin{bmatrix} a & b \\ cn & d \end{bmatrix} \mid a, b, c, d \in \mathbb{Z},\ ad - bcn = 1 \right\} \tag{2.5.2}$$

According to [CN79] the normalizer $N(\Gamma_0(n))$ of $\Gamma_0(n)$ in $G(\mathbb{R})$ is commensurable with $G(\mathbb{Z})$ and admits the description

$$N(\Gamma_0(n)) = \left\{ \begin{bmatrix} ae & b/h \\ cn/h & de \end{bmatrix} \mid a, b, c, d \in \mathbb{Z},\ e \in \mathbb{Z}^+,\ e \| n/h,\ ade - bcn/eh^2 = 1 \right\} \tag{2.5.3}$$

where $h = ((n, 24)$ is the largest divisor of 24 such that h^2 divides n (cf. §2.2). The expression (2.5.3) tells us, in particular, that $N(\Gamma_0(n))_\infty = \langle T^{1/h} \rangle$ for $h = ((n, 24)$.

Proposition 2.5.1. *Let $n \in \mathbb{Z}^+$. Then $N(\Gamma_0(n))$ acts transitively on $\hat{\mathbb{Q}}$. We have $N(\Gamma_0(n))_\infty = \Gamma_0(n)_\infty$ if and only if n is not divisible by 4 or 9.*

An important family of groups, each one commensurable with the modular group $G(\mathbb{Z})$, and each one containing and normalizing some $\Gamma_0(n)$, was introduced in [CN79]; these are the groups of $n\|h$*-type*, and we now recall their definition. For $n, h \in \mathbb{Z}^+$ with $h|n$, and for S a subgroup of $\mathrm{Ex}(n/h)$ (cf. §2.2), we define a group $\Gamma_0(n|h) + S$ by setting

$$\Gamma_0(n|h) + S = \left\{ \begin{bmatrix} ae & b/h \\ cn & de \end{bmatrix} \mid a, b, c, d \in \mathbb{Z},\ e \in S,\ ade - bcn/eh = 1 \right\}. \tag{2.5.4}$$

Evidently the group $\Gamma_0(n|h)+S$ contains $\Gamma_0(nh)$. In case $h|24$ it also normalizes $\Gamma_0(nh)$. Assume then that $h|24$. We say that $\Gamma_0(n|h) + S$ is of *Fricke type* if the element $\left[\begin{smallmatrix} 0 & -1 \\ nh & 0 \end{smallmatrix}\right]$ belongs to it.

The group $\Gamma_0(n\|h)+S$ is defined (cf. [CN79]) as the subgroup of $\Gamma_0(n|h)+S$ arising as the kernel of a certain morphism $\Gamma_0(n|h)+S\to\mathbb{Z}/h$ which factors through the canonical map $\Gamma_0(n|h)+S\to(\Gamma_0(n|h)+S)/\Gamma_0(nh)$. In order to describe it let S' denote the image of S in $\mathrm{Ex}(nh)$ under the natural injection $\mathrm{Ex}(n/h)\to\mathrm{Ex}(nh)$ (cf. §2.2). Then as generators for the quotient $(\Gamma_0(n|h)+S)/\Gamma_0(nh)$ we may take the cosets X, Y, and $W_{e'}$ for $e'\in S'$, given by

$$X=\begin{bmatrix}1 & 1/h\\ 0 & 1\end{bmatrix}\Gamma_0(nh),\quad Y=\begin{bmatrix}1 & 0\\ n & 1\end{bmatrix}\Gamma_0(nh),\quad W_{e'}=\begin{bmatrix}ae' & b\\ cnh & de'\end{bmatrix}\Gamma_0(nh),\tag{2.5.5}$$

where the a, b, c and d in the definition of $W_{e'}$ are arbitrary integers for which $ade'-bcnh/e'=1$, and we may define $\Gamma_0(n\|h)+S$ to be the kernel of the composition

$$\Gamma_0(n|h)+S\to(\Gamma_0(n|h)+S)/\Gamma_0(nh)\xrightarrow{\lambda}\mathbb{Z}/h\tag{2.5.6}$$

where the map λ is determined by the requirements that $\lambda(W_{e'})=0$ for all $e'\in S'$, $\lambda(X)=1$, and $\lambda(Y)$ is 1 or -1 according as $\Gamma_0(n|h)+S$ is of Fricke type or not.

The group $\Gamma_0(n\|h)+S$ is also described in [Fer93], where it is denoted $(1/h)\Gamma_0(n|h)+e_1,e_2,\ldots$ for $S=\{1,e_1,e_2,\ldots\}$, although we should remark that certain variations on λ are considered there (in the case that h is even), so that the class of groups considered in [Fer93] properly contains the groups of $n\|h$-type.

2.6 Scalings

Let Γ be a group commensurable with $G(\mathbb{Z})$ and let $\mathsf{p}\in\mathsf{P}_\Gamma$ (cf. (2.5.1)) be a cusp of Γ. An element $p\in\mathsf{p}\subset\hat{\mathbb{Q}}$ will be called a *representative for* p.

Lemma 2.6.1. *Let Γ be a group commensurable with $G(\mathbb{Z})$ and let $p\in\hat{\mathbb{Q}}$. Then there exists an element $\sigma_p\in G(\mathbb{Q})$ such that*

$$p=\sigma_p\cdot\infty,\quad(\sigma_p^{-1}\Gamma\sigma_p)_\infty=B(\mathbb{Z}).\tag{2.6.1}$$

Proof. In case $p=\infty$ the group Γ_∞ (cf. (2.3.7)) of elements in Γ that fix ∞ is infinite cyclic and generated by some element $\gamma_\infty\in B(\mathbb{Q})$. We claim that in fact γ_∞ lies in $B_u(\mathbb{Q})$, and thus equals T^α (cf. (2.3.4)) for some $\alpha\in\mathbb{Q}^+$. For otherwise $P\mathrm{det}(\gamma_\infty)=e$ for some $e>1$ (cf. (2.3.6)), so that $P\mathrm{det}(\gamma_\infty^n)=e^n$ for $n\in\mathbb{Z}$, since the restriction $P\mathrm{det}:B(\mathbb{Q})\to\mathbb{Z}^+$ is multiplicative. Then each power of γ_∞ lies in a distinct coset of the intersection $\Gamma\cap G(\mathbb{Z})$ in Γ since $P\mathrm{det}$ is invariant under multiplication by elements of $G(\mathbb{Z})$. This contradicts the hypothesis that Γ is commensurable with $G(\mathbb{Z})$, so we conclude that $\Gamma_\infty=\langle T^\alpha\rangle$ for some $\alpha\in\mathbb{Q}^+$. Then we may

take $\sigma_\infty = [\alpha]$ (cf. (2.3.3)), for upon calculating $[\mu]T^\alpha[1/\mu] = T^{\mu\alpha}$ we find that

$$([1/\alpha]\Gamma[\alpha])_\infty = [1/\alpha]\Gamma_\infty[\alpha] = [1/\alpha]\langle T^\alpha\rangle[\alpha] = \langle T\rangle, \tag{2.6.2}$$

and the group $\langle T\rangle$ is just $B(\mathbb{Z})$, so the element $\sigma_\infty = [\alpha]$ satisfies the two conditions (2.6.1).

In case $p = a/c$ for coprime integers $a, c \in \mathbb{Z}$ with $c \neq 0$, we may choose $b, d \in \mathbb{Z}$ such that $ad - bc = 1$ and set $\sigma = \left[\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right]$. Then $p = \sigma \cdot \infty$ and σ satisfies the first condition in (2.6.1). Now we rerun the argument of the paragraph above with $\sigma^{-1}\Gamma\sigma$ in place of Γ to find that $(\sigma^{-1}\Gamma\sigma)_\infty = \langle T^\alpha\rangle$ for some $\alpha \in \mathbb{Q}^+$, and $\sigma_p = \sigma[\alpha]$ satisfies the two conditions (2.6.1). This completes the proof of the lemma. □

An element $\sigma_p \in G(\mathbb{Q})$ satisfying the two conditions (2.6.1) of Lemma 2.6.1 will be called a *scaling element for* Γ *at the cusp representative* p. It is useful to have a replacement for the notion of scaling element that is independent of a choice of cusp representative.

Lemma 2.6.2. *Let* Γ *be a group commensurable with* $G(\mathbb{Z})$ *and let* $\mathsf{p} \in \mathsf{P}_\Gamma$ *be a cusp of* Γ. *Then there exists a coset* $\Sigma_\mathsf{p} \in \Gamma\backslash G(\mathbb{Q})$ *such that*

$$\mathsf{p} = \Sigma_\mathsf{p} \cdot \infty, \quad (\Sigma_\mathsf{p}^{-1}\Sigma_\mathsf{p})_\infty = B(\mathbb{Z}). \tag{2.6.3}$$

Proof. Let $p \in \mathsf{p} \subset \hat{\mathbb{Q}}$ be a representative for p and let $\sigma_p \in G(\mathbb{Q})$ satisfy the conditions (2.6.1) of Lemma 2.6.1. Then the coset $\Sigma_\mathsf{p} = \Gamma\sigma_p$ satisfies the conditions (2.6.3). □

A coset $\Sigma_\mathsf{p} \in \Gamma\backslash G(\mathbb{Q})$ satisfying the two conditions (2.6.3) of Lemma 2.6.2 will be called a *scaling coset for* Γ *at the cusp* p. Any element $\sigma \in \Sigma_\mathsf{p}$ is a scaling element for Γ at the representative $\sigma \cdot \infty$ for the cusp p, and if σ_p is a scaling element for Γ at the cusp representative p then the coset $\Gamma\sigma_p$ is a scaling coset for Γ at the cusp $\mathsf{p} = \Gamma \cdot p$ represented by p.

Scaling cosets are determined only up to right multiplication by elements of $B_u(\mathbb{Q})$. Indeed, both the conditions defining the notion of scaling coset are invariant under the replacement of Σ_p by $\Sigma_\mathsf{p}T^\alpha$ for some $\alpha \in \mathbb{Q}$ (cf. (2.3.4)). A set $\{\Sigma_\mathsf{p} \mid \mathsf{p} \in \mathsf{P}_\Gamma\} \subset \Gamma\backslash G(\mathbb{Q})$ such that Σ_p is a scaling coset for Γ at p for each $\mathsf{p} \in \mathsf{P}_\Gamma$ will be called a *system of scaling cosets for* Γ.

The problem of constructing scaling cosets for a group Γ may be viewed in the following way. The group $G(\mathbb{Q})$ acts transitively on $\hat{\mathbb{Q}}$, and the subgroup $B(\mathbb{Q})$ is just the stabilizer of the distinguished point $\infty \in \hat{\mathbb{Q}}$. Thus the map $\Gamma\backslash G(\mathbb{Q}) \to \mathsf{P}_\Gamma$ given by $\Gamma\sigma \mapsto \Gamma\sigma \cdot \infty$ induces an isomorphism $\Gamma\backslash G(\mathbb{Q})/B(\mathbb{Q}) \cong \mathsf{P}_\Gamma$ which we may regard as identifying the sets $\Gamma\backslash G(\mathbb{Q})/B(\mathbb{Q})$ and P_Γ. Now $B(\mathbb{Q})$ is naturally isomorphic to the semidirect product $B_u(\mathbb{Q}) \rtimes B_d(\mathbb{Q})$ (cf. §2.5), so the natural map $\Gamma\backslash G(\mathbb{Q}) \to \Gamma\backslash G(\mathbb{Q})/B(\mathbb{Q})$ factors through $\Gamma\backslash G(\mathbb{Q})/B_u(\mathbb{Q})$, yielding a

naturally defined sequence

$$\Gamma\backslash G(\mathbb{Q}) \to \Gamma\backslash G(\mathbb{Q})/B_u(\mathbb{Q}) \to \Gamma\backslash G(\mathbb{Q})/B(\mathbb{Q}) \cong \mathsf{P}_\Gamma \tag{2.6.4}$$

where the fibres of the second map are torsors for the diagonal group $B_d(\mathbb{Q}) \simeq \mathbb{Q}^+$. Given a cusp $\mathsf{p} \in \mathsf{P}_\Gamma$, any preimage of p in $\Gamma\backslash G(\mathbb{Q})$ under the composition (2.6.4) is a coset of Γ satisfying the first condition defining a scaling coset for Γ at p. In order to satisfy also the second condition we should multiply this coset by $[\mu]$ (cf. (2.3.3)) for some (uniquely defined) $\mu \in \mathbb{Q}^+$. Thus the two conditions defining scaling cosets define a section of the $B_d(\mathbb{Q})$-bundle $\Gamma\backslash G(\mathbb{Q})/B_u(\mathbb{Q}) \to \mathsf{P}_\Gamma$. We denote this map $\mathsf{p} \mapsto \mathfrak{S}_\mathsf{p}$. We may naturally identify $\mathfrak{S}_\mathsf{p}$ with the set of scaling cosets for Γ at p, for these are precisely the preimages of $\mathfrak{S}_\mathsf{p}$ in $\Gamma\backslash G(\mathbb{Q})$ under the first map of (2.6.4).

$$\begin{aligned} \mathsf{P}_\Gamma &\to \Gamma\backslash G(\mathbb{Q})/B_u(\mathbb{Q}) \\ \mathsf{p} &\mapsto \mathfrak{S}_\mathsf{p} = \{\Sigma_\mathsf{p} T^\alpha \mid \alpha \in \mathbb{Q}\} \end{aligned} \tag{2.6.5}$$

One says that Γ has *width one at infinity* in the case that $\Gamma_\infty = B(\mathbb{Z})$. Observe that Γ has width one at infinity if and only if Γ is a scaling coset for itself at the infinite cusp $\Gamma\cdot\infty$; that is, if and only if $\Gamma \in \mathfrak{S}_{\Gamma\cdot\infty}$. If Γ does not have width one at infinity then there is a unique $\mu \in \mathbb{Q}^+$ with the property that $\Gamma[\mu]$ is a scaling coset for Γ at the infinite cusp $\Gamma\cdot\infty$, and then the group $\Gamma^{[\mu]} = [1/\mu]\Gamma[\mu]$ is a group with width one at infinity.

When engaged in the task of computing Fourier coefficients of modular forms for a group Γ say, one frequently has use for double coset spaces of the form $\Gamma_p\backslash\Gamma/\Gamma_q$, for some $p, q \in \hat{\mathbb{Q}}$, where Γ_p denotes the stabilizer in Γ of p. If σ_p and σ_q are scaling elements for Γ at p and q, respectively, then we have $\sigma_p^{-1}\Gamma_p\sigma_p = B(\mathbb{Z})$, and similarly with q in place of p, so the set of *translates*

$$\sigma_p^{-1}(\Gamma_p\backslash\Gamma/\Gamma_q)\sigma_q = \{\sigma_p^{-1}\Gamma_p\gamma\Gamma_q\sigma_q \mid \gamma \in \Gamma\} \tag{2.6.6}$$

is in fact a set of double cosets of $B(\mathbb{Z})$. In the notation of §2.4 we have

$$\sigma_p^{-1}(\Gamma_p\backslash\Gamma/\Gamma_q)\sigma_q = \left[\!\left[\sigma_p^{-1}\Gamma\sigma_q\right]\!\right] = \left\{\left[\!\left[\sigma_p^{-1}\gamma\sigma_q\right]\!\right] \mid \gamma \in \Gamma\right\}. \tag{2.6.7}$$

Set $\mathsf{p} = \Gamma\cdot p$ and $\mathsf{q} = \Gamma\cdot q$, and set $\Sigma_\mathsf{p} = \Gamma\sigma_p$ and $\Sigma_\mathsf{q} = \Gamma\sigma_q$, so that Σ_p and Σ_q are scaling cosets for Γ at p and q, respectively. Then we have $\sigma_p^{-1}\Gamma\sigma_q = \Sigma_\mathsf{p}^{-1}\Sigma_\mathsf{q}$, so that these translates depend only on cusps, and not on cusp representatives. We see from this discussion that for any pair of cusps $\mathsf{p}, \mathsf{q} \in \mathsf{P}_\Gamma$, with scaling cosets Σ_p and Σ_q, respectively, the set $\Sigma_\mathsf{p}^{-1}\Sigma_\mathsf{q}$ is a union of double cosets of $B(\mathbb{Z})$. The assignment $(\mathsf{p}, \mathsf{q}) \mapsto \Sigma_\mathsf{p}^{-1}\Sigma_\mathsf{q}$ is sensitive to the choice of scaling cosets Σ_p and Σ_q, to the extent that a different choice will replace $\Sigma_\mathsf{p}^{-1}\Sigma_\mathsf{q}$ with a set of the form $T^\alpha\Sigma_\mathsf{p}^{-1}\Sigma_\mathsf{q}T^\beta$ for some $\alpha, \beta \in \mathbb{Q}$.

In general there may be no canonical choice of scaling coset for a particular group at a particular cusp, but there are situations in which some choices might be preferred over others. For example, it is natural to take $\Sigma_{\Gamma\cdot\infty} = \Gamma$ in case Γ has width one at infinity, for we then have $\Sigma_{\mathsf{p}}^{-1}\Sigma_{\mathsf{q}} = \Sigma_{\mathsf{p}}^{-1}$ when $\mathsf{q} = \Gamma\cdot\infty$, and $\Sigma_{\mathsf{p}}^{-1}\Sigma_{\mathsf{q}} = \Sigma_{\mathsf{q}}$ when $\mathsf{p} = \Gamma\cdot\infty$, and $\Sigma_{\mathsf{p}}^{-1}\Sigma_{\mathsf{q}} = \Gamma$ when $\mathsf{p} = \mathsf{q} = \Gamma\cdot\infty$. We shall always take the scaling coset $\Sigma_{\Gamma\cdot\infty}$, for Γ at the infinite cusp, to be of the form $\Gamma[\mu]$ for $\mu \in \mathbb{Q}^+$. This value μ is uniquely determined.

Observe that if Σ_{p} and Σ_{q} are scaling cosets for Γ at p and q, respectively, then the intersection $\Sigma_{\mathsf{p}}^{-1}\Sigma_{\mathsf{q}} \cap B(\mathbb{Q}) = (\Sigma_{\mathsf{p}}^{-1}\Sigma_{\mathsf{q}})_\infty$ can be non-empty only in the case that $\mathsf{p} = \mathsf{q}$. For p and q cusps for Γ we define $\delta_{\Gamma|\mathsf{p},\mathsf{q}}$ to be 1 or 0 according as $\mathsf{p} = \mathsf{q}$ or not.

Given a system $\{\Sigma_{\mathsf{p}} \mid \mathsf{p} \in \mathsf{P}_\Gamma\}$ of scaling cosets for Γ, we write Γ^{p} as a shorthand for the group obtained as the conjugate of Γ by Σ_{p}.

$$\Gamma^{\mathsf{p}} = \Sigma_{\mathsf{p}}^{-1}\Sigma_{\mathsf{p}} \tag{2.6.8}$$

The notation (2.6.8) suppresses the dependence on the choice of scaling coset: a different choice will replace Γ^{p} with a group of the form $T^{-\alpha}\Gamma^{\mathsf{p}}T^{\alpha}$ for some $\alpha \in \mathbb{Q}$.

2.7 Integrals

Write $\mathcal{O}(\mathbb{H})$ for the ring of holomorphic functions on the upper half plane $\mathbb{H}$.

For $\kappa \in \mathbb{Z}$, we define the *weight* 2κ *(right) action of* $G(\mathbb{Q})$ *on* $\mathcal{O}(\mathbb{H})$, to be denoted $(f,\chi) \mapsto f|^\kappa\chi$, by setting

$$(f|^\kappa\chi)(z) = f(\chi\cdot z)\,\mathrm{jac}(\chi,z)^\kappa \tag{2.7.1}$$

for $f \in \mathcal{O}(\mathbb{H})$ and $\chi \in G(\mathbb{Q})$. For Γ a group commensurable with $G(\mathbb{Z})$ (and hence a subgroup of $G(\mathbb{Q})$), we call $f \in \mathcal{O}(\mathbb{H})$ an *unrestricted modular form of weight* 2κ *for* Γ in case it is a fixed point for the weight 2κ action of Γ. Suppose f is an unrestricted modular form of weight 2κ for Γ. Then for $X \in \Gamma\backslash G(\mathbb{Q})$ a right coset of Γ in $G(\mathbb{Q})$ we may define a function $(f\|_\Gamma^\kappa X)(z)$ by setting

$$f\|_\Gamma^\kappa X = f|^\kappa\chi \tag{2.7.2}$$

where χ is any representative for the coset $X \in \Gamma\backslash G(\mathbb{Q})$. Let $\mathsf{p} \in \mathsf{P}_\Gamma$ be a cusp of Γ and let $\Sigma_{\mathsf{p}} \in \Gamma\backslash G(\mathbb{Q})$ a scaling coset for Γ at p. Then for f an unrestricted modular form of weight 2κ for Γ we define $f_{|\mathsf{p}} \in \mathcal{O}(\mathbb{H})$ by setting $f_{|\mathsf{p}} = f\|_\Gamma^\kappa\Sigma_{\mathsf{p}}$. Then $f_{|\mathsf{p}}(z+1) = f_{|\mathsf{p}}(z)$ for all $z \in \mathbb{H}$, by

the defining properties of Σ_{p} (cf. §2.6), so we have

$$f_{|\mathsf{p}}(z) = \sum_{n\in\mathbb{Z}} c_{|\mathsf{p}}(n)\,\mathrm{e}(nz) \tag{2.7.3}$$

for some $c_{|\mathsf{p}}(n) \in \mathbb{Z}$. We call the right hand side of (2.7.3) the *Fourier expansion of f at p with respect to Σ_{p}*, and we say that f is *meromorphic at* p if the right hand side of (2.7.3) is a meromorphic function of $q = \mathrm{e}(z)$ in a neighborhood of $q = 0$. A different choice of scaling coset will replace $c_{|\mathsf{p}}(n)$ with $\mathrm{e}(n\alpha)c_{|\mathsf{p}}(n)$ for some $\alpha \in \mathbb{Q}$, so the notion of being meromorphic at p is independent of the choice of scaling coset at p.

We define $M_\kappa(\Gamma)$ to be the vector space consisting of unrestricted modular forms of weight 2κ for Γ that are meromorphic at the cusps of Γ. In a slight departure from standard convention we call $M_\kappa(\Gamma)$ the space of *modular forms of weight* 2κ *for* Γ. We define $S_\kappa(\Gamma)$ to be the vector space consisting of modular forms of weight 2κ for Γ that vanish at the cusps of Γ. We call $S_\kappa(\Gamma)$ the space of *cusp forms of weight* 2κ *for* Γ. The vector space $S_\kappa(\Gamma)$ is the zero vector space when $\kappa \leq 0$.

Let $\kappa \in \mathbb{Z}$ such that $\kappa \leq 0$, and let $w \in \mathbb{H} \cup \hat{\mathbb{Q}}$. Given a holomorphic function $g(z)$ on $\mathbb{H}$ with sufficiently rapid decay as the imaginary part of z tends to ∞ we define a function $I_w^\kappa g$ on $\mathbb{C}$ by setting $I_\infty^\kappa g$ to be the zero function and by setting

$$(I_w^\kappa g)(z) = 2\pi\mathbf{i}\int_w^\infty g(\xi)(2\pi\mathbf{i}(\xi - z))^{(-2\kappa)}\mathrm{d}\xi \tag{2.7.4}$$

for $w \neq \infty$, where the integral is taken over the vertical line $\{w + \mathbf{i}t \mid t \in \mathbb{R}^+\}$. In our applications g will either be a cusp form (for some $\Gamma < PSL_2(\mathbb{R})$ commensurable with $PSL_2(\mathbb{Z})$) or will admit a finite power series expansion of the form $g(z) = \sum_{n=1}^N c(n)\,\mathrm{e}(nz)$. In these cases the convergence of the integral defining $I_w^\kappa g$ is clear.

We define a closely related operator $g \mapsto J_w^\kappa g$ by setting

$$(J_w^\kappa g)(z) = \overline{(I_w^\kappa g)(\bar{z})}. \tag{2.7.5}$$

Observe that when $\kappa \leq 0$, the functions $(I_w^\kappa g)(z)$ and $(J_w^\kappa g)(z)$ are polynomials of degree at most -2κ in z . For $\chi \in G(\mathbb{Q})$ we have

$$(I_w^\kappa g)|^\kappa\,\chi = I_{w\cdot\chi}^\kappa\left(g|^{1-\kappa}\chi\right) - I_{\infty\cdot\chi}^\kappa\left(g|^{1-\kappa}\chi\right) \tag{2.7.6}$$

where $w \cdot \chi = \chi^{-1} \cdot w$ denotes the natural right action of $G(\mathbb{Q})$ on $\mathbb{H} \cup \hat{\mathbb{Q}}$. In particular, if $g \in M_{1-\kappa}(\Gamma)$ for some group Γ then we have $(I_w^\kappa g)|^\kappa\,\gamma = I_{w\cdot\gamma}^\kappa g - I_{\infty\cdot\gamma}^\kappa g$ for $\gamma \in \Gamma$.

Lemma 2.7.1. *Let Γ be a group commensurable with $G(\mathbb{Z})$, let $\kappa \in \mathbb{Z}$ such that $\kappa \leq 0$, and let $g \in S_{1-\kappa}(\Gamma)$. Then we have $I_p^\kappa g = 0$ for all $p \in \hat{\mathbb{Q}}$ if and only if $g = 0$.*

Proof. Certainly $I_p^\kappa g = 0$ for all $p \in \hat{\mathbb{Q}}$ if $g = 0$, so let $g \in S_{1-\kappa}(\Gamma)$ such that $I_p^\kappa g = 0$ for all $p \in \hat{\mathbb{Q}}$. Following the proof of Lemma 3.2 in [Nie74] we define a function $f(z)$ by setting

$$f(z) = (I_z^\kappa g)(z) = 2\pi\mathbf{i}\int_z^\infty g(\xi)(2\pi\mathbf{i}(\xi - z))^{(-2\kappa)}\mathrm{d}\xi. \tag{2.7.7}$$

The identity (2.7.6) implies $f|^\kappa\gamma = f - I_{\infty\cdot\gamma}^\kappa g$ for $\gamma \in \Gamma$, so that f is an unrestricted modular form of weight 2κ for Γ. We should consider the behavior of f at the cusps of Γ. Let $\mathsf{p} \in \mathsf{P}_\Gamma$, and let Σ_p be a scaling coset for Γ at p. Then we have

$$g_{|\mathsf{p}}(\xi) = \sum_{n>0} b_{|\mathsf{p}}(n)\,\mathrm{e}(n\xi) \tag{2.7.8}$$

for some $b_{|\mathsf{p}}(n) \in \mathbb{C}$, where $g_{|\mathsf{p}} = g\|_\Gamma^{1-\kappa}\Sigma_\mathsf{p}$ (cf. (2.7.2)). In order to analyze $f_{|\mathsf{p}} = f\|_\Gamma^\kappa\Sigma_\mathsf{p}$ we choose $\sigma \in \Sigma_\mathsf{p}$ and compute

$$f|^\kappa\sigma = I_z^\kappa(g_{|\mathsf{p}}) - I_{\infty\cdot\sigma}^\kappa(g_{|\mathsf{p}}) \tag{2.7.9}$$

by (2.7.6). Now the term $I_{\infty\cdot\sigma}^\kappa(g_{|\mathsf{p}})$ vanishes, since we have

$$\begin{aligned}(I_{\infty\cdot\sigma}^\kappa(g_{|\mathsf{p}}))(z) &= 2\pi\mathbf{i}^{1-2\kappa}\int_{\sigma^{-1}\cdot\infty}^\infty g(\sigma\cdot\xi)\,\mathrm{jac}(\sigma,\xi)^{1-\kappa}(\xi - z)^{(-2\kappa)}\mathrm{d}\xi \\ &= 2\pi\mathbf{i}^{1-2\kappa}\int_{\sigma^{-1}\cdot\infty}^\infty g(\sigma\cdot\xi)(\sigma\cdot\xi - \sigma\cdot z)^{(-2\kappa)}\,\mathrm{jac}(\sigma,z)^\kappa\mathrm{d}(\sigma\cdot\xi) \\ &= -(I_{\sigma\cdot\infty}^\kappa g)\,\mathrm{jac}(\sigma,z)^\kappa,\end{aligned} \tag{2.7.10}$$

which vanishes by our hypothesis on g. So we have the Fourier expansion

$$f_{|\mathsf{p}}(z) = \sum_{n>0} c_{|\mathsf{p}}(n)\,\mathrm{e}(nz) \tag{2.7.11}$$

for f at p, where $c_{|\mathsf{p}}(n)$ is determined by $b_{|\mathsf{p}}(n)$ according to the formula

$$c_{|\mathsf{p}}(n) = b_{|\mathsf{p}}(n)2\pi\mathbf{i}(I_0^\kappa g)(0) = -n^{2\kappa-1}b_\mathsf{p}(n). \tag{2.7.12}$$

We conclude that f is a cusp form of weight κ. Since $\kappa \leq 0$ we have $S_\kappa(\Gamma) = 0$, and so $c_{|\mathsf{p}}(n) = 0$ for all $n \in \mathbb{Z}$ and all $\mathsf{p} \in \mathsf{P}_\Gamma$, and so $b_{|\mathsf{p}}(n) = 0$ for all $n \in \mathbb{Z}$ and all $\mathsf{p} \in \mathsf{P}_\Gamma$. That is to say, g vanishes identically, which is what we required to show. □

Let $\kappa \in \mathbb{Z}$ such that $\kappa \leq 0$. Following Niebur (cf. [Nie74]) we say that a holomorphic function $f \in \mathcal{O}(\mathbb{H})$ is an *unrestricted automorphic integral of weight* 2κ *for* Γ in case there exists

a cusp form $g \in S_{1-\kappa}(\Gamma)$ such that

$$f|^\kappa \gamma = f - J^\kappa_{\infty\cdot\gamma} g \tag{2.7.13}$$

for each $\gamma \in \Gamma$. According to Lemma 2.7.1 the cusp form $g \in S_{1-\kappa}(\Gamma)$ satisfying (2.7.13) is uniquely determined. We call it the *cusp form associated to the unrestricted automorphic integral* f. The operator $f \mapsto f\|^\kappa_\Gamma X$, for X a right coset of Γ, may be extended to unrestricted automorphic integrals as follows. Suppose f is an unrestricted automorphic integral of weight 2κ for Γ with associated cusp form g. Then for $X \in \Gamma\backslash G(\mathbb{Q})$ we may define $f\|^\kappa_\Gamma X \in \mathcal{O}(\mathbb{H})$ by setting

$$f\|^\kappa_\Gamma X = (f - J^\kappa_{\chi\cdot\infty} g)|^\kappa \chi \tag{2.7.14}$$

where χ is any representative for the coset $X \in \Gamma\backslash G(\mathbb{Q})$. The transformation properties (2.7.6) and (2.7.13) confirm that the function $f\|^\kappa_\Gamma X$ is independent of the choice of coset representative χ. Let Σ_{p} be a scaling coset for Γ at a cusp $\mathsf{p} \in \mathsf{P}_\Gamma$, and define $f_{|\mathsf{p}}$ by setting

$$f_{|\mathsf{p}} = f\|^\kappa_\Gamma \Sigma_{\mathsf{p}}. \tag{2.7.15}$$

By the defining properties of scaling cosets we find that $f_{|\mathsf{p}}(z+1) = f_{|\mathsf{p}}(z)$ for all $z \in \mathbb{H}$, and so we have

$$f_{|\mathsf{p}}(z) = \sum_{n\in\mathbb{Z}} c_{|\mathsf{p}}(n)\, \mathrm{e}(nz) \tag{2.7.16}$$

for some $c_{|\mathsf{p}}(n) \in \mathbb{C}$. We call (2.7.16) the *Fourier expansion of* f *at* p *with respect to* Σ_{p}, and we say that f is *meromorphic at* p if the right hand side of (2.7.16) is a meromorphic function of $q = \mathrm{e}(z)$ in a neighborhood of $q = 0$.

We define $I_\kappa(\Gamma)$ to be the vector space consisting of unrestricted automorphic integrals of weight 2κ for Γ that are meromorphic at every cusp of Γ. We call $I_\kappa(\Gamma)$ the space of *automorphic integrals of weight* 2κ *for* Γ.

The space $M_\kappa(\Gamma)$ of modular forms of weight 2κ for Γ is a subspace of $I_\kappa(\Gamma)$ by definition. Lemma 2.7.1 may be regarded as stating that $M_\kappa(\Gamma)$ is the kernel of the map $I_\kappa(\Gamma) \to S_{1-\kappa}(\Gamma)$ which sends an automorphic integral f to its associated cusp form g (cf. (2.7.13)). We will see in §4.5 that the map $I_\kappa(\Gamma) \to S_{1-\kappa}(\Gamma)$ is surjective, so that we have an exact sequence of vector spaces

$$0 \to M_\kappa(\Gamma) \to I_\kappa(\Gamma) \to S_{1-\kappa}(\Gamma) \to 0. \tag{2.7.17}$$

The sequence (2.7.17) is trivially exact in case $\kappa > 0$, for then $S_{1-\kappa}(\Gamma) = 0$, so that an automorphic integral of weight 2κ is automatically a modular form. Indeed, and more generally, if $\kappa \in \mathbb{Z}$ is chosen so that $S_{1-\kappa}(\Gamma) = 0$, then the spaces $M_\kappa(\Gamma)$ and $I_\kappa(\Gamma)$ coincide by the definition (cf. (2.7.13)) of an automorphic integral.

3 Rademacher sums

In this section we associate Rademacher sums, and conjugate Rademacher sums (cf. §1.3), of arbitrary even integer weight to triples $(\Gamma, \mathfrak{p}, \mathfrak{q})$ where Γ is a group commensurable with the modular group and $\mathfrak{p}$ and $\mathfrak{q}$ are cusps for Γ. We derive explicit expressions for the Fourier expansions of these functions, and study how they transform under the action of Γ.

3.1 Construction

Given $\kappa, m \in \mathbb{Z}$ and $[\![\chi \in [\![G(\mathbb{Q})$, define a function $z \mapsto P_{[\![\chi}^{\kappa(m)}(z)$ on $\mathbb{H}$ by setting

$$P_{[\![\chi}^{\kappa(m)}(z) = \mathrm{e}(m\,[\![\chi \cdot z)\,\mathrm{jac}([\![\chi\,, z)^\kappa. \tag{3.1.1}$$

Given $U \subset [\![G(\mathbb{Q})$, define the *holomorphic Poincaré series of weight* 2κ *and order* m *associated to* U, to be denoted $P_U^{\kappa(m)}(z)$, by setting

$$P_U^{\kappa(m)}(z) = \lim_{K\to\infty} \sum_{[\![\chi \in U_{\leq K}} P_{[\![\chi}^{\kappa(m)}(z) \tag{3.1.2}$$

where $U_{\leq K}$ is as in (2.4.6). Of course, the limit defining $P_U^{\kappa(m)}(z)$ may or may not converge depending on the choice of κ, m, and U. If $\kappa > 1$ and $m \geq 0$ then for $U = [\![\Gamma$ say, for Γ a group commensurable with $G(\mathbb{Z})$, the limiting sum in (3.1.2) converges absolutely and uniformly on compact subsets of $\mathbb{H}$ (cf. Lemma 3.4.2) and the symbols $\lim_{K\to\infty}$ and $_{\leq K}$ are unnecessary. If $\kappa \leq 0$ then even with the limit the expression on the right hand side of (3.1.2) fails to converge, and the Poincaré series $P_U^{\kappa(m)}(z)$ is not defined.

As was originally demonstrated by Rademacher in [Rad39], the non-convergence of the Poincaré series $P_U^{\kappa(m)}(z)$ when $\kappa = 0$ can be circumvented by replacing $P_{[\![\chi}^{0(m)}(z)$ with $P_{[\![\chi}^{0(m)}(z) - P_{0,[\![\chi}^{0(m)}$ in the right-hand side of (3.1.2) where

$$P_{[\![\chi}^{0(m)}(z) - P_{0,[\![\chi}^{0(m)} = \mathrm{e}(m\,[\![\chi \cdot z) - \mathrm{e}(m\,[\![\chi \cdot \infty) \tag{3.1.3}$$

in case $[\![\chi \in [\![G(\mathbb{Q})^\times$, and $P_{0,[\![\chi}^{0(m)} = 0$ otherwise. Rademacher showed for $m = -1$ (cf. loc. cit.) that the resulting expression is a (conditionally) convergent series which recovers a function invariant for the modular group $G(\mathbb{Z})$ in the case that $U = [\![G(\mathbb{Z})$. Generalizations of

Rademacher's construction were given by Knopp (cf. [Kno62a]) and Niebur (cf. [Nie74]) so as to obtain conditionally convergent Poincaré series for arbitrary real non-positive weights, for various subgroups of $G(\mathbb{R}) = PSL_2(\mathbb{R})$.

With the constructions of Rademacher, Knopp and Niebur in mind, we introduce the *Rademacher component function of weight* 2κ *and order* m *associated to* $[\![\chi$, denoted $z \mapsto R_{[\![\chi}^{\kappa(m)}(z)$ and defined for $\kappa, m \in \mathbb{Z}$ and $[\![\chi \in [\![G(\mathbb{Q})$, by setting

$$R_{[\![\chi}^{\kappa(m)}(z) = \mathrm{e}(m\,[\![\chi \cdot z)\,\mathrm{r}^{\kappa}(m, [\![\chi\,, z)\,\mathrm{jac}([\![\chi\,, z)^{\kappa}, \tag{3.1.4}$$

where $\mathrm{r}^{\kappa}(m, [\![\chi\,, z)$ is the *Rademacher regularization factor of weight* 2κ, given by

$$\mathrm{r}^{\kappa}(m, [\![\chi\,, z) = \Phi(1-2\kappa, 2-2\kappa, m\,[\![\chi \cdot \infty - m\,[\![\chi \cdot z)(2\pi\mathbf{i}(m\,[\![\chi \cdot z - m\,[\![\chi \cdot \infty))^{1-2\kappa} \tag{3.1.5}$$

in case $\chi \in G(\mathbb{Q})^{\times}$, and $\mathrm{r}^{\kappa}(m, [\![\chi\,, z) = 1$ otherwise (cf. (2.1.3)). The function $R_{[\![\chi}^{\kappa(m)}(z)$ is entire in case $\chi \in B(\mathbb{Q})$, and is otherwise holomorphic away from $z = \chi^{-1} \cdot \infty$. For $U \subset [\![G(\mathbb{Q})$ we define the *classical Rademacher sum of weight* 2κ *and order* m *associated to* U, to be denoted $R_U^{\kappa(m)}(z)$, by setting

$$R_U^{\kappa(m)}(z) = \lim_{K \to \infty} \sum_{[\![\chi \in U_{\leq K}} R_{[\![\chi}^{\kappa(m)}(z). \tag{3.1.6}$$

Again, the expression defining $R_U^{\kappa(m)}(z)$ will not converge for all choices of κ, m and U. Employing the Kummer transformation (2.1.4) we may rewrite the Rademacher component function $R_{[\![\chi}^{\kappa(m)}(z)$ as

$$R_{[\![\chi}^{\kappa(m)}(z) = \mathrm{e}(m\,[\![\chi \cdot \infty)\,\mathrm{e}(m\,[\![\chi \cdot z - m\,[\![\chi \cdot \infty, 1-2\kappa)\,\mathrm{jac}([\![\chi\,, z)^{\kappa} \tag{3.1.7}$$

in case $[\![\chi \in [\![G(\mathbb{Q})^{\times}$, where $\mathrm{e}(z,s)$ is the generalized exponential function of (2.1.6). Since we restrict attention to the case that κ is an integer (cf. [Nie74] for non-integral weights), the generalized exponential in (3.1.7) may in turn be written in terms of the partial exponential function $\mathrm{e}(z)_{<K}$ (cf. §2.1). We thus obtain the identity

$$\mathrm{e}(m\,[\![\chi \cdot z)\,\mathrm{r}^{\kappa}(m, [\![\chi\,, z) = \mathrm{e}(m\,[\![\chi \cdot z) - \mathrm{e}(m\,[\![\chi \cdot \infty)\,\mathrm{e}(m\,[\![\chi \cdot z - m\,[\![\chi \cdot \infty)_{<1-2\kappa} \tag{3.1.8}$$

when $\chi \in G(\mathbb{Q})^{\times}$. In particular, we have $\mathrm{r}^{\kappa}(m, [\![\chi\,, z) = 1$ when κ is positive, so that the Rademacher sum $R_U^{\kappa(m)}(z)$ is exactly (albeit formally) the Poincaré series $P_U^{\kappa(m)}(z)$ when $\kappa > 0$.

The case that $\kappa = 0$ is special, for the corresponding Poincaré series, if it were to converge, would define a Γ-invariant function on $\mathbb{H}$ for any group Γ satisfying $U\Gamma = U$. With $\kappa = 0$ we

have $\mathrm{r}^0(m, [\![\chi\,, z) = 1 - \mathrm{e}(m\,[\![\chi \cdot \infty - m\,[\![\chi \cdot z)$ when $\chi \in G(\mathbb{Q})^\times$, and hence

$$R^{0(m)}_{[\![\chi}(z) = P^{0(m)}_{[\![\chi}(z) - P^{0(m)}_{0,[\![\chi} \tag{3.1.9}$$

for all $[\![\chi \in [\![G(\mathbb{Q})^\times$. In other words, the Rademacher sum $R^{0(m)}_U(z)$ is given by

$$\begin{aligned} R^{0(m)}_U(z) &= \lim_{K\to\infty} \sum_{[\![\chi \in U^\times_{\leq K}} P^{0(m)}_{[\![\chi}(z) - P^{0(m)}_{0,[\![\chi} \\ &= \sum_{[\![\chi \in U_\infty} \mathrm{e}(m\,[\![\chi \cdot z) + \lim_{K\to\infty} \sum_{[\![\chi \in U^\times_{\leq K}} \mathrm{e}(m\,[\![\chi \cdot z) - \mathrm{e}(m\,[\![\chi \cdot \infty), \end{aligned} \tag{3.1.10}$$

with the latter expression holding at least when there are only finitely many cosets of $B(\mathbb{Z})$ in U_∞. This expression (3.1.10), with $m = -1$ and $U = [\![G(\mathbb{Z})\,$, is the series originally considered by Rademacher in [Rad39].

It is interesting to consider the function obtained by substituting $\bar{z}$ for z in the right hand side of (3.1.6). We define the *conjugate Rademacher sum of weight* 2κ *and order* m *associated to* U, to be denoted $\bar{R}^{\kappa(m)}_U(z)$, by setting

$$\bar{R}^{\kappa(m)}_U(z) = \lim_{K\to\infty} \sum_{[\![\chi \in U_{\leq K}} \bar{R}^{\kappa(m)}_{[\![\chi}(z), \tag{3.1.11}$$

where $\bar{R}^{\kappa(m)}_{[\![\chi}(z) = R^{\kappa(m)}_{[\![\chi}(\bar{z})$ by definition. Evidently, the conjugate Rademacher sum $\bar{R}^{\kappa(m)}_U(z)$ is an anti-holomorphic function on $\mathbb{H}$ whenever U, κ and m are such that the limit in (3.1.11) converges locally uniformly in z. At first glance it appears that we should recover the classical Rademacher sum $R^{\kappa(m)}_U(z)$ from the conjugate Rademacher sum according to the identity $R^{\kappa(m)}_U(z) = \bar{R}^{\kappa(m)}_U(\bar{z})$, but as we shall see in §3.4 the relationship between the holomorphic functions $R^{\kappa(m)}_U(z)$ and $\bar{R}^{\kappa(m)}_U(\bar{z})$ is generally more interesting that this.

We typically take U to be of the form $U = [\![\Sigma^{-1}_{\mathsf{p}}\Sigma_{\mathsf{q}}$ where $\{\Sigma_{\mathsf{p}} \mid \mathsf{p} \in \mathsf{P}_\Gamma\}$ is a system of scaling cosets (cf. §2.6) for some group Γ commensurable with the modular group, and $\mathsf{p}, \mathsf{q} \in \mathsf{P}_\Gamma$ are cusps of Γ. In this case we write $R^{\kappa(m)}_{\Gamma,\mathsf{p}|\mathsf{q}}(z)$ for $R^{\kappa(m)}_U(z)$, and similarly for the holomorphic Poincaré series, suppressing the choice of scaling cosets from notation. A change in the choice of Σ_{p} and Σ_{q} replaces $R^{\kappa(m)}_{\Gamma,\mathsf{p}|\mathsf{q}}(z)$ with a function of the form $\mathrm{e}(\alpha)R^{\kappa(m)}_{\Gamma,\mathsf{p}|\mathsf{q}}(z+\beta)$ for some $\alpha, \beta \in \mathbb{Q}$, and similarly for the holomorphic Poincaré series $P^{\kappa(m)}_{\Gamma,\mathsf{p}|\mathsf{q}}(z)$.

In the case that p or q is the *infinite cusp* $\Gamma \cdot \infty$ we omit it from notation, writing $R^{\kappa(m)}_{\Gamma,\mathsf{p}}(z)$ for $R^{\kappa(m)}_{\Gamma,\mathsf{p}|\Gamma\cdot\infty}(z)$, and $R^{\kappa(m)}_{\Gamma|\mathsf{q}}(z)$ for $R^{\kappa(m)}_{\Gamma,\Gamma\cdot\infty|\mathsf{q}}(z)$, and similarly for the holomorphic Poincaré series. The functions $R^{\kappa(m)}_{\Gamma,\mathsf{p}}(z)$ are, in a sense, the most important, for we shall see in §3.4 that $R^{\kappa(m)}_{\Gamma,\mathsf{p}}(z)$ is, up to addition by a certain constant function, an automorphic integral of weight 2κ for Γ with a single pole at the cusp p in case Γ has width one at infinity (cf. §2.6) and $\kappa \leq 0$ and $m < 0$. We shall see also in §3.4 (cf. Theorems 3.4.3 and 3.4.4) that the function $R^{\kappa(m)}_{\Gamma,\mathsf{p}|\mathsf{q}}(z)$,

once corrected by adding a certain constant function in case $\kappa \leq 0$, is the expansion (of the correction by addition of a certain constant) of $R^{\kappa(m)}_{\Gamma,\mathfrak{p}}(z)$ at the cusp $\mathfrak{q}$ in the sense of §2.7.

To further emphasize the importance of the functions $R^{\kappa(m)}_{\Gamma,\mathfrak{p}}(z)$ we observe that every Rademacher sum $R^{\kappa(m)}_{\Gamma,\mathfrak{p}|\mathfrak{q}}(z)$ is of the form $R^{\kappa(m)}_{\Gamma',\mathfrak{p}'}(z)$ for some group Γ' with width one at infinity, and some cusp $\mathfrak{p}'$ of Γ'. For if $\Sigma_{\mathfrak{p}}$ and $\Sigma_{\mathfrak{q}}$ are the chosen scaling sets for Γ at $\mathfrak{p}$ and $\mathfrak{q}$, respectively, then we have $R^{\kappa(m)}_{\Gamma,\mathfrak{p}|\mathfrak{q}}(z) = R^{\kappa(m)}_{U}(z)$ for $U = [\![\Sigma_{\mathfrak{p}}^{-1}\Sigma_{\mathfrak{q}}$. Recall (cf. (2.6.8)) that $\Gamma^{\mathfrak{q}}$ is a shorthand for the group $\Sigma_{\mathfrak{q}}^{-1}\Sigma_{\mathfrak{q}}$, and this group $\Gamma^{\mathfrak{q}}$ has width one at infinity by the defining properties of scaling cosets. If we define $\mathfrak{p}^{\mathfrak{q}} = \Sigma_{\mathfrak{q}}^{-1} \cdot \mathfrak{p} \subset \hat{\mathbb{Q}}$ then $\mathfrak{p}^{\mathfrak{q}}$ is a cusp of $\Gamma^{\mathfrak{q}}$, and $\Sigma_{\mathfrak{p}^{\mathfrak{q}}} = \Sigma_{\mathfrak{q}}^{-1}\Sigma_{\mathfrak{p}}$ is a scaling coset for $\Gamma^{\mathfrak{q}}$ at $\mathfrak{p}^{\mathfrak{q}}$. We now have $\Sigma_{\mathfrak{p}}^{-1}\Sigma_{\mathfrak{q}} = \Sigma_{\mathfrak{p}^{\mathfrak{q}}}^{-1}$, so that $U = [\![\Sigma_{\mathfrak{p}^{\mathfrak{q}}}^{-1}$. Since $\Gamma^{\mathfrak{q}}$ has width one at infinity we may take it to be a scaling coset for itself at the infinite cusp, and we thus have

$$R^{\kappa(m)}_{\Gamma,\mathfrak{p}|\mathfrak{q}}(z) = R^{\kappa(m)}_{\Gamma^{\mathfrak{q}},\mathfrak{p}^{\mathfrak{q}}}(z) \tag{3.1.12}$$

subject to a consistent choice of scaling cosets for Γ and $\Gamma^{\mathfrak{q}}$. More explicitly, the equality in (3.1.12) holds in case $\Gamma^{\mathfrak{q}} = \Sigma_{\mathfrak{q}}^{-1}\Sigma_{\mathfrak{q}}$ and $\mathfrak{p}^{\mathfrak{q}} = \Sigma_{\mathfrak{q}}^{-1} \cdot \mathfrak{p}$ and $\Sigma_{\mathfrak{p}^{\mathfrak{q}}} = \Sigma_{\mathfrak{q}}^{-1}\Sigma_{\mathfrak{p}}$.

Our primary interest in this article is in the distinguished case that $\kappa = 0$. In order to simplify notation, and maintain consistency with the notation of §1, we write $R^{(m)}_{\Gamma,\mathfrak{p}|\mathfrak{q}}(z)$ as a shorthand for $R^{0(m)}_{\Gamma,\mathfrak{p}|\mathfrak{q}}(z)$.

We conclude this section with a result which expresses the Rademacher component function $R^{\kappa(m)}_{[\![\chi}(z)$ in terms of the functions $P^{\kappa(m)}_{[\![\chi}(z)$ of (3.1.1) and the integral operator J^{κ}_{w} of §2.7.

Lemma 3.1.1. *Let $\kappa, m \in \mathbb{Z}$ such that $\kappa \leq 0$ and $m < 0$, and let $[\![\chi \in [\![G(\mathbb{Q})$. Then we have*

$$R^{\kappa(m)}_{[\![\chi}(z) = P^{\kappa(m)}_{[\![\chi}(z) - m^{1-2\kappa}\left(J^{\kappa}_{\infty\cdot[\![\chi}P^{1-\kappa(-m)}_{[\![\chi}\right)(z). \tag{3.1.13}$$

Proof. In case $[\![\chi \in [\![G(\mathbb{Q})_{\infty}$ the second term in the right hand side of (3.1.13) vanishes (cf. (3.1.15)), and the identity (3.1.13) then holds by the definition of $R^{\kappa(m)}_{[\![\chi}(z)$, so we assume henceforth that $[\![\chi \in [\![G(\mathbb{Q})^{\times}$. By (3.1.8) we may write the Rademacher component function $R^{\kappa(m)}_{[\![\chi}(z)$ as

$$\mathrm{e}(m\,[\![\chi\cdot z)\,\mathrm{jac}([\![\chi\,,z)^{\kappa} - \mathrm{e}(m\,[\![\chi\cdot\infty)\,\mathrm{e}(m\,[\![\chi\cdot z - m\,[\![\chi\cdot\infty)_{<1-2\kappa}\,\mathrm{jac}([\![\chi\,,z)^{\kappa}. \tag{3.1.14}$$

The first term in (3.1.14) is just $P^{\kappa(m)}_{[\![\chi}(z)$, and a contour integral calculation confirms that the second term in (3.1.14) is $m^{1-2\kappa}$ times the image of $P^{1-\kappa(-m)}_{[\![\chi}(z)$ under $J^{\kappa}_{\infty\cdot[\![\chi}$. The details of this contour integral calculation are as follows. We have

$$\overline{\left(J^{\kappa}_{\infty\cdot[\![\chi}P^{1-\kappa(-m)}_{[\![\chi}\right)(z)} = (2\pi\mathbf{i})^{1-2\kappa}\int_{\chi^{-1}]\cdot\infty}^{\infty}\mathrm{e}(-m\,[\![\chi\cdot\xi)\,\mathrm{jac}([\![\chi\,,\xi)^{1-\kappa}(\xi-\bar{z})^{(-2\kappa)}\mathrm{d}\xi \tag{3.1.15}$$

by the definition of J_w^κ (cf. §2.7). Choosing a representative χ for the coset $[\![\chi$ we compute

$$\begin{aligned}
&\int_{\chi^{-1}]\!]\cdot\infty}^{\infty} \mathrm{e}(-m\,[\![\chi\cdot\xi)\,\mathrm{jac}([\![\chi\,,\xi)^{1-\kappa}(\xi-\bar{z})^{(-2\kappa)}\mathrm{d}\xi \\
&= \mathrm{e}(-m\chi\cdot\infty)\int_{\infty}^{\chi\cdot\infty} \mathrm{e}(m\chi\cdot\infty - m\chi\cdot\xi)(\chi\cdot\xi-\chi\cdot\bar{z})^{(-2\kappa)}\mathrm{d}(\chi\cdot\xi)\,\mathrm{jac}(\chi,\bar{z})^{\kappa} \\
&= -\,\mathrm{e}(-m\chi\cdot\infty)\int_{\chi\cdot\infty}^{\infty} \mathrm{e}(m\chi\cdot\infty - m\xi)(\xi-\chi\cdot\bar{z})^{(-2\kappa)}\mathrm{d}\xi\,\mathrm{jac}(\chi,\bar{z})^{\kappa} \\
&= -\,\mathrm{e}(-m\chi\cdot\infty)\int_{0}^{\infty} \mathrm{e}(-m\xi)(\xi+\chi\cdot\infty-\chi\cdot\bar{z})^{(-2\kappa)}\mathrm{d}\xi\,\mathrm{jac}(\chi,\bar{z})^{\kappa}
\end{aligned} \tag{3.1.16}$$

where the last integral in (3.1.16) is taken over the vertical line $\{\mathbf{i}t \mid t\in\mathbb{R}^+\}$. Thus, making the substitution $2\pi\mathbf{i}m\xi = t$, we find that

$$\begin{aligned}
&\int_{\chi^{-1}]\!]\cdot\infty}^{\infty} \mathrm{e}(-m\,[\![\chi\cdot\xi)\,\mathrm{jac}([\![\chi\,,\xi)^{1-\kappa}(\xi-\bar{z})^{(-2\kappa)}\mathrm{d}\xi \\
&= -\,\mathrm{e}(-m\chi\cdot\infty)(2\pi\mathbf{i}m)^{-1}\int_{0}^{\infty} e^{-t}((2\pi\mathbf{i}m)^{-1}t+\chi\cdot\infty-\chi\cdot\bar{z})^{(-2\kappa)}\mathrm{d}t\,\mathrm{jac}(\chi,\bar{z})^{\kappa} \\
&= -\,\mathrm{e}(-m\chi\cdot\infty)\sum_{k=0}^{-2\kappa}(2\pi\mathbf{i}m)^{2\kappa-1}(2\pi\mathbf{i}(m\chi\cdot\infty-m\chi\cdot\bar{z}))^{(k)}\,\mathrm{jac}(\chi,\bar{z})^{\kappa},
\end{aligned} \tag{3.1.17}$$

and from this we deduce that

$$\left(J^{\kappa}_{\infty\cdot[\![\chi}\,P^{1-\kappa(-m)}_{[\![\chi}\right)(z) = m^{2\kappa-1}\,\mathrm{e}(m\chi\cdot\infty)\,\mathrm{e}(m\chi\cdot z - m\chi\cdot\infty)_{<1-2\kappa}\,\mathrm{jac}(\chi,z)^{\kappa}. \tag{3.1.18}$$

Upon comparison with (3.1.14) we obtain the required identity (3.1.13). This completes the proof. □

3.2 Coefficients

For $[\![\chi]\!]\in[\![G(\mathbb{Q})]\!]^\times$ and $m,n\in\mathbb{Z}$ we define a holomorphic function $s\mapsto \mathrm{Kl}^s_{[\![\chi]\!]}(m,n)$ on $\mathbb{C}$ by setting

$$\mathrm{Kl}^s_{[\![\chi]\!]}(m,n) = \mathrm{e}(m\,[\![\chi]\!]\cdot\infty)\,\mathrm{e}(-n\,[\![\chi^{-1}]\!]\cdot\infty)\varrho\,[\![\chi]\!]^s\,. \tag{3.2.1}$$

Here $\varrho\,[\![\chi]\!]$ denotes the scaling factor of (2.3.11) (cf. Lemma 2.4.1). If $\chi\in G(\mathbb{Q})^\times$ has a preferred representative $\left(\begin{smallmatrix}a & b\\ c & d\end{smallmatrix}\right)$ with $c>0$, then we have

$$\mathrm{Kl}^s_{[\![\chi]\!]}(m,n) = \mathrm{e}\left(\frac{ma+nd}{c}\right)\frac{(ad-bc)^s}{c^{2s}}, \tag{3.2.2}$$

so that sums of the $\mathrm{Kl}^s_{[\![\chi]\!]}(m,n)$ over suitable subsets of $[\![G(\mathbb{Q})]\!]^\times$ recover zeta functions of various kinds, as we shall see presently. With this in mind consider the series $\mathrm{Kl}^s_S(m,n)$ defined for S a subset of $[\![G(\mathbb{Q})]\!]^\times$ by setting

$$\mathrm{Kl}^s_S(m,n) = \lim_{K\to\infty} \sum_{[\![\chi]\!]\in S_{\leq K}} \mathrm{Kl}^s_{[\![\chi]\!]}(m,n) \tag{3.2.3}$$

where $S_{\leq K}$ is as in (2.4.7). Write $\mathrm{Kl}^s_{\Gamma,\mathsf{p}|\mathsf{q}}(m,n)$ for $\mathrm{Kl}^s_S(m,n)$ in the case that $S = [\![\Sigma_\mathsf{p}^{-1}\Sigma_\mathsf{q}]\!]^\times$ for Γ a subgroup of $G(\mathbb{Q})$ that is commensurable with $G(\mathbb{Z})$ and $\{\Sigma_\mathsf{p} \mid \mathsf{p}\in\mathsf{P}_\Gamma\}$ is a system of scaling cosets for Γ. We will verify presently (cf. Proposition 3.2.1) that the Dirichlet series $s\mapsto \mathrm{Kl}^s_{\Gamma,\mathsf{p}|\mathsf{q}}(m,n)$ converges absolutely for $\Re(s)>1$ (so that the limit is unnecessary there) and therefore defines a holomorphic function in this region; these functions $\mathrm{Kl}^s_{\Gamma,\mathsf{p}|\mathsf{q}}(m,n)$ are among the *Selberg–Kloosterman zeta functions* introduced by Selberg in [Sel65] and shown there to admit meromorphic continuation to the entire complex plane. Selberg demonstrated further in loc. cit. that the eigenfunctions of the hyperbolic Laplacian $\Delta = (z-\bar z)^2\partial_z\partial_{\bar z}$ on $\Gamma\backslash\mathbb{H}$ determine the poles of $\mathrm{Kl}^s_S(m,n)$ in such a way that a pole at s corresponds to an eigenfunction with eigenvalue $\lambda = s(1-s)$. The convergence of expressions very similar to (3.2.3) in certain regions $\Re(s)\geq\sigma$ with $\sigma\leq 1$ was considered by Knopp in [Kno86] and [Kno89].[1]

Proposition 3.2.1. *Let $\Gamma < G(\mathbb{Q})$ be a group commensurable with $G(\mathbb{Z})$, let $\{\Sigma_\mathsf{p} \mid \mathsf{p}\in\mathsf{P}_\Gamma\}$ be a system of scaling cosets for Γ and set $S = [\![\Sigma_\mathsf{p}^{-1}\Sigma_\mathsf{q}]\!]^\times$ for some $\mathsf{p},\mathsf{q}\in\mathsf{P}_\Gamma$. Then the series (3.2.3) defining $\mathrm{Kl}^s_S(m,n)$ converges for $s=1$ when m and n are not both zero, and converges absolutely and locally uniformly in $\Re(s)>1$ for any $m,n\in\mathbb{Z}$.*

Proof. For positive integers r and c let S^r_c denote the subset of S consisting of double cosets $[\![\chi]\!]\in[\![G(\mathbb{Q})]\!]^\times$ with $P\det(\chi)=r$ and $c(\chi)=c$, so that $\varrho(\chi)=r/c^2$ (cf. (2.3.6)) for $[\![\chi]\!]\in S^r_c$. Observe that there are at most c elements in S^r_c so we have

$$|\mathrm{Kl}^s_{S^r_c}(m,n)| \leq \sum_{[\![\chi]\!]\in S^r_c} |\mathrm{Kl}^s_{[\![\chi]\!]}(m,n)| \leq r^s c^{1-2s} \tag{3.2.4}$$

for $s\in\mathbb{C}$. Now $S=\bigcup_{r,c>0} S^r_c$ and there are only finitely many r for which $S^r=\bigcup_{c>0}S^r_c$ is non-empty since Γ is commensurable with $G(\mathbb{Z})$, so (3.2.4) and the identity $\mathrm{Kl}^s_S(m,n) = \sum_{c>0}\sum_{r>0}\mathrm{Kl}^s_{S^r_c}(m,n)$ implies that the series of absolute values obtained from (3.2.3) is bounded above by a constant independent of m and n so long as $\Re(s)>1$. Indeed, we have

$$|\mathrm{Kl}^s_S(m,n)| \leq \sum_{c>0}\sum_{r\in P\det(S)} r^\sigma c^{1-2\sigma} \leq \#P\det(S) r_{\max}^\sigma \frac{1}{2\sigma-2} \tag{3.2.5}$$

[1] Please note that the latter of these references contains important corrections to the former.

where $\#P\mathrm{det}(S)$ is the number of values of r for which S^r is non-empty, $r_{\max}$ is the maximum r for which S^r is non-empty, and $\sigma = \Re(s)$. This proves the required convergence of (3.2.3) in the region $\Re(s) > 1$.

We now consider convergence at $s = 1$. As mentioned above, Selberg demonstrated the meromorphic continuation of the Selberg–Kloosterman zeta function $s \to \mathrm{Kl}_S^s(m,n)$ and related its poles to the spectrum of the Laplacian on $\Gamma\backslash\mathbb{H}$ in [Sel65]. It turns out that those poles in $\Re(s) > 1/2$ are finite in number and are confined to the real line segment $(1/2, 1)$ in the case that m and n are not both zero, and the absence of a pole at $s = 1$ derives from the fact that the eigenfunctions for Δ with eigenvalue 0 are harmonic, and thus constant. (See [Iwa02], and Chapter 9 especially, for a thorough development of these ideas.) In particular, if we write $Z_{m,n}(s)$ for Selberg's meromorphic continuation of $\mathrm{Kl}_S^s(m,n)$, then $Z_{m,n}(s)$ is holomorphic at s when $\Re(s) = 1$ and m and n are not both zero, and the only pole of $Z_{0,0}(s)$ on the line $\Re(s) = 1$ is at $s = 1$. On the other hand Goldfeld–Sarnak determined the estimate

$$|Z_{m,n}(s)| = O\left(|m||n|\frac{|s|^{1/2}}{\Re(s) - 1/2}\right) \tag{3.2.6}$$

for $mn \neq 0$ as $\Im(s) \to \infty$ with $\Re(s) > 1/2$ in [GS83]. (Loc. cit. considers the case that $m, n > 0$. See [Pri00] for the extension to $mn \neq 0$, and an improvement upon (3.2.6).) Armed with this bound and the holomorphy of $Z_{m,n}(s)$ in $\Re(s) > 1 - a$ for some $a > 0$ we can deduce the convergence of (3.2.3) at $s = 1$ in the following way, which is an adaptation of the argument presented in §2 of [Kno86]. Note that we can choose $a > 0$ such that $Z_{m,n}(s)$ has no poles in $(1-a, 1)$ for all $m, n \in \mathbb{Z}$ since the values of s such that $\Re(s) > 1/2$ and $s(1-s)$ is an eigenvalue of Δ are finite in number. For such a the estimate (3.2.6) holds uniformly in s for $\Re(s) > 1 - a$.

Let $a > 0$ such that the interval $(1 - a, 1)$ contains no poles of $Z_{m,n}(s)$ for any $m, n \in \mathbb{Z}$, let $T > 0$ and set C to be the positively oriented rectangular contour with corners $\pm a \pm \mathbf{i}T$. Consider the contour integral

$$I = \frac{1}{2\pi\mathbf{i}} \int_C Z_{m,n}\left(\frac{t}{2} + 1\right) x^t \frac{\mathrm{d}t}{t} \tag{3.2.7}$$

where $x > 0$. By our choice of a the integrand has a unique simple pole inside C and the residue there is $Z_{m,n}(1)$, so $I = Z_{m,n}(1)$. On the other hand the integral over the right most vertical segment of C contributes $\Sigma_T(x)$ where

$$\Sigma_T(x) = \frac{1}{2\pi\mathbf{i}} \int_{a-\mathbf{i}T}^{a+\mathbf{i}T} Z_{m,n}\left(\frac{t}{2} + 1\right) x^t \frac{\mathrm{d}t}{t}. \tag{3.2.8}$$

Observe that in the limit as $T \to \infty$ we obtain the function

$$\Sigma(x) = \lim_{T\to\infty} \Sigma_T(x) = \sum_{r>0} \sum_{0<c<x\sqrt{r}} \mathrm{Kl}^1_{S_c^r}(m,n) \tag{3.2.9}$$

(so long as x does not coincide with $c/\sqrt{r}$ for some $c, r > 0$ with S_c^r non-empty) and the difference $|\Sigma(x) - \Sigma_T(x)|$ is $O(x^a T^{-1})$ where the implied constant is independent of m and n. The estimate (3.2.6) implies that the contribution to I of the left most vertical segment of C is $O(|m||n|x^{-a}T^{1/2})$, and we may show that the contributions of the two horizontal components of C are $O(|m||n|x^a T^{1/4-1/4a})$ so long as $T < x^{4a}$ by applying the Phragmén–Lindelöf theorem (as stated in Theorem 14 of [HR64]) to the estimates $|Z_{m,n}(t/2+1)| = O(|m||n||\Im(t)|^{1/2})$ for $\Re(t) = -a$ and $|Z_{m,n}(t/2+1)| = O(1)$ for $\Re(t) = a$ (where the former of these is another application of (3.2.6) and the constant implied by the latter is independent of both m and n according to Proposition 3.2.1). Setting $T = x^{3a/2}$ we obtain

$$|\Sigma(x) - Z_{m,n}(1)| = O(|m||n|x^{-a/4}) \tag{3.2.10}$$

under the condition that $a < 3/13$, and in the limit as $x \to \infty$ we arrive at the convergence of the sum (3.2.3) at $s = 1$ in the case that $mn \neq 0$.

For the case that $mn = 0$ but $m + n \neq 0$ we note that a similar estimate to (3.2.6)

$$|Z_{m,n}(s)| = O\left(|s|^{1/2}\right) \tag{3.2.11}$$

holds uniformly in s for $\Re(s) > 1 - a$, with the implied constant independent of m and n, as a consequence of the boundedness of Fourier coefficients of non-analytic Eisenstein series in this region (cf. [Kub73]). Now the argument of the preceding paragraph goes through with (3.2.11) used in place of (3.2.6) and we obtain the desired convergence of $\mathrm{Kl}^s_S(m,n)$ at $s = 1$ for all $m, n \in \mathbb{Z}$ such that m and n are not both zero. This completes the proof of the proposition. □

Remark 3.2.2. Consider the case that $S = [\![G(\mathbb{Z})]\!]^\times$. Then the function $\mathrm{Kl}^s_S(m,n)$ is closely related to the Riemann zeta function $\zeta(s)$, for we have

$$\mathrm{Kl}^s_S(m,0) = \sum_{c\in\mathbb{Z}^+} \sum_{a\in(\mathbb{Z}/c)^*} \frac{\mathrm{e}(ma/c)}{c^{2s}} = \frac{\sigma(|m|, 1-2s)}{\zeta(2s)} \tag{3.2.12}$$

for $\Re(s) > 1$ in the case that $m < 0$, where $\sigma(n,s)$ is the divisor function of (2.2.1). In particular, the function $s \mapsto \mathrm{Kl}^{s/2}_S(1,0)$ is just the reciprocal of the Riemann zeta function when $S = [\![G(\mathbb{Z})]\!]^\times$. When $m = n = 0$ we have $\mathrm{Kl}^s_S(0,0) = \sum_{c>0} \phi(c) c^{-2s}$ where ϕ is Euler's totient function so clearly the sum diverges at $s = 1$ in this case. In general the residue of $\mathrm{Kl}^s_S(0,0)$ at $s = 1$ is the reciprocal of the hyperbolic area of the quotient $\Gamma\backslash\mathbb{H}$ when $S = [\![\Gamma]\!]^\times$ for some

group Γ. (Cf. [Iwa02, §6.4].)

The following result is evident from the definition of $\mathrm{Kl}^s_{[\![\chi]\!]}(m,n)$, and the definition of the conjugation $\chi \mapsto \bar{\chi}$ (cf. §2.3).

Lemma 3.2.3. *For any $\chi \in G(\mathbb{Q})^\times$, $m,n \in \mathbb{Z}$, and $s \in \mathbb{C}$, we have*

$$\mathrm{Kl}^s_{[\![\chi]\!]}(m,n) = \mathrm{Kl}^s_{[\![\chi^{-1}]\!]}(n,m) = \mathrm{Kl}^s_{[\![\bar{\chi}]\!]}(-m,-n), \tag{3.2.13}$$

$$\mathrm{Kl}^s_{[\![\chi]\!]}(m,n)\varrho\,[\![\chi]\!]^{1-s} = \mathrm{Kl}^{1-s}_{[\![\chi]\!]}(m,n)\varrho\,[\![\chi]\!]^{s}\,. \tag{3.2.14}$$

For a continuous analogue of the function $\mathrm{Kl}^s_{[\![\chi]\!]}(m,n)$ we define $\mathrm{Bl}^\kappa_{[\![\chi]\!]}(w,z)$, for $\chi \in G(\mathbb{Q})^\times$, $w,z \in \mathbb{C}$ and $\kappa \in \mathbb{Z}$, by setting

$$\mathrm{Bl}^\kappa_{[\![\chi]\!]}(w,z) = 2\pi\mathbf{i}\,\mathrm{Res}_{\xi=0}\,\mathrm{e}\left(\frac{w\varrho\,[\![\chi]\!]}{\xi}\right)\mathrm{e}(z\xi)\,\frac{1}{\xi^{2\kappa}}\,\mathrm{d}\xi. \tag{3.2.15}$$

Lemma 3.2.4. *The function $\mathrm{Bl}^\kappa_{[\![\chi]\!]}(w,z)$ admits the following series representations.*

$$\mathrm{Bl}^\kappa_{[\![\chi]\!]}(w,z) = \sum_{k\geq 0}(2\pi\mathbf{i})^{2k+2\kappa}\varrho\,[\![\chi]\!]^k\,w^{(k)}z^{(k+2\kappa-1)} \tag{3.2.16}$$

$$\mathrm{Bl}^\kappa_{[\![\chi]\!]}(w,z) = \sum_{k\geq 0}(2\pi\mathbf{i})^{2k+2-2\kappa}\varrho\,[\![\chi]\!]^{k+1-2\kappa}\,w^{(k+1-2\kappa)}z^{(k)} \tag{3.2.17}$$

The first expression (3.2.16) holds for $\kappa > 0$. The second expression (3.2.17) holds for $\kappa \leq 0$.

Lemma 3.2.5. *For any $\chi \in G(\mathbb{Q})^\times$ and $w,z \in \mathbb{C}$ and $\kappa \in \mathbb{Z}$ we have*

$$\mathrm{Bl}^\kappa_{[\![\chi]\!]}(w,z)\varrho\,[\![\chi]\!]^{\kappa} = -\,\mathrm{Bl}^{1-\kappa}_{[\![\chi]\!]}(z,w)\varrho\,[\![\chi]\!]^{1-\kappa} = -\,\mathrm{Bl}^\kappa_{[\![\chi]\!]}(-w,-z)\varrho\,[\![\chi]\!]^{\kappa}\,. \tag{3.2.18}$$

The main construction of this section is the following. Given a subset $S \subset [\![G(\mathbb{Q})]\!]^\times$ and $\kappa \in \mathbb{Z}$, we define the *coefficient function*, denoted $(m,n) \mapsto \mathrm{c}^\kappa_S(m,n)$, by setting

$$\mathrm{c}^\kappa_S(m,n) = \lim_{K\to\infty}\sum_{[\![\chi]\!]\in S_{\leq K}}\mathrm{Kl}^\kappa_{[\![\chi]\!]}(m,n)\,\mathrm{Bl}^\kappa_{[\![\chi]\!]}(m,n) \tag{3.2.19}$$

(cf. (2.4.7)) whenever this limit converges. In applications we take S to be of the form $S = [\![\Sigma_{\mathsf{p}}^{-1}\Sigma_{\mathsf{q}}]\!]^\times$ where Σ_{p} and Σ_{q} are scaling cosets for a group Γ at cusps $\mathsf{p},\mathsf{q} \in \mathsf{P}_\Gamma$ where Γ is commensurable with $G(\mathbb{Z})$. We will see presently that $\mathrm{c}^\kappa_S(m,n)$ converges for any such S. We obtain the following formulas by combining (3.2.1) and Lemma 3.2.4.

Lemma 3.2.6. *The coefficient functions $\mathrm{c}_S^\kappa(m,n)$ admit the following series representations.*

$$\mathrm{c}_S^\kappa(m,n) = \lim_{K\to\infty} \sum_{[\![\chi]\!]\in S_{\le K}} \sum_{k\ge 0} \mathrm{e}(m\,[\![\chi]\!]\cdot\infty)\,\mathrm{e}(-n\,[\![\chi^{-1}]\!]\cdot\infty)(-4\pi^2\varrho\,[\![\chi]\!])^{k+\kappa} m^{(k)} n^{(k+2\kappa-1)} \tag{3.2.20}$$

$$\mathrm{c}_S^\kappa(m,n) = \lim_{K\to\infty} \sum_{[\![\chi]\!]\in S_{\le K}} \sum_{k\ge 0} \mathrm{e}(m\,[\![\chi]\!]\cdot\infty)\,\mathrm{e}(-n\,[\![\chi^{-1}]\!]\cdot\infty)(-4\pi^2\varrho\,[\![\chi]\!])^{k+1-\kappa} m^{(k+1-2\kappa)} n^{(k)} \tag{3.2.21}$$

The expression (3.2.20) holds for $\kappa > 0$. The expression (3.2.21) holds for $\kappa \le 0$.

Proposition 3.2.7. *For $S \subset [\![G(\mathbb{Q})]\!]^\times$ and $m,n,\kappa \in \mathbb{Z}$ we have*

$$\mathrm{c}_S^\kappa(m,n) = -\mathrm{c}_{S^{-1}}^{1-\kappa}(n,m) = -\mathrm{c}_{\bar S}^\kappa(-m,-n). \tag{3.2.22}$$

Proof. These identities follow directly from Lemma 3.2.6, and the observation that the values $\varrho\,[\![\chi]\!]$, $\varrho\,[\![\chi^{-1}]\!]$ and $\varrho\,[\![\bar\chi]\!]$ all coincide for any $\chi \in G(\mathbb{Q})^\times$. □

Proposition 3.2.8. *Let Γ be a group commensurable with $G(\mathbb{Z})$ and let $\mathsf{p},\mathsf{q} \in \mathsf{P}_\Gamma$ be cusps of Γ. Then the expression (3.2.19) defining the coefficient function $\mathrm{c}_S^\kappa(m,n)$ converges for $S = [\![\Sigma_{\mathsf{p}}^{-1}\Sigma_{\mathsf{q}}]\!]^\times$ and any $\kappa,m,n \in \mathbb{Z}$, and further, we have*

$$|\mathrm{c}_S^\kappa(m,n)| = O\left(|m|^{1/2-\kappa}|n|^{\kappa-1/2} e^{C|mn|^{1/2}}\right) \tag{3.2.23}$$

as $|n| \to \infty$ where C depends only on S and the implied constant depends only on S and κ.

Proof. Observe that according to the $\kappa \mapsto 1-\kappa$ symmetry of Proposition 3.2.7 it suffices to consider the case that $\kappa > 0$. Observe also that for $\kappa > 0$ the coefficient function $\mathrm{c}_S^\kappa(m,n)$ vanishes for all m when $n = 0$, and when $m = 0$ we have

$$\mathrm{c}_S^\kappa(0,n) = (-4\pi^2)^\kappa n^{(2\kappa-1)}\,\mathrm{Kl}_S^\kappa(0,n) \tag{3.2.24}$$

according to (3.2.20), so the convergence in case $mn = 0$ is either trivial or follows from the convergence of the Selberg–Kloosterman zeta function $\mathrm{Kl}_S^\kappa(m,n)$ which is established in Proposition 3.2.1. Assume then that $mn \ne 0$. By (3.2.20) we have

$$\mathrm{c}_S^\kappa(m,n) = (-4\pi^2)^\kappa \sum_{c>0}\sum_{r>0}\sum_{k\ge 0} \mathrm{Kl}_{S_c^r}^{\kappa+k}(m,n)(-4\pi^2 m)^{(k)} n^{(k+2\kappa-1)} \tag{3.2.25}$$

where S_c^r is defined as in the proof of Proposition 3.2.1 and the sum over r is finite. In case

$\kappa > 1$ we can reorder the summation over k in (3.2.25) past the other two since the sum

$$\mathrm{c}_S^\kappa(m,n) = (-4\pi^2)^\kappa \sum_{k\geq 0} (-4\pi^2 m)^{(k)} n^{(k+2\kappa-1)}\, \mathrm{Kl}_S^{\kappa+k}(m,n) \tag{3.2.26}$$

so obtained is absolutely convergent according to the estimate (3.2.5). In detail,

$$\begin{aligned} |\mathrm{c}_S^\kappa(m,n)| &\leq (4\pi^2)^\kappa \sum_{k\geq 0} (4\pi^2|m|)^{(k)} |n|^{(k+2\kappa-1)} \#P\det(S) r_{\max}^{\kappa+k} \frac{1}{2\kappa-2+2k} \\ &\leq \#P\det(S)(4\pi^2 r_{\max})^\kappa \sum_{k\geq 0} (4\pi^2 r_{\max}|m|)^{(k)} |n|^{(k+2\kappa-1)} < \infty. \end{aligned} \tag{3.2.27}$$

Now for A and B non-negative we have

$$\sum_{k\geq 0} A^{(k)} B^{(k+2\kappa-1)} \leq A^{1/2-\kappa} B^{\kappa-1/2} \sum_{k\geq 0} (2(AB)^{1/2})^{(2k+2\kappa-1)} \tag{3.2.28}$$

so that setting $A = 4\pi^2 r_{\max}|m|$ and $B = |n|$ in (3.2.28) we obtain the required estimate (3.2.23) from (3.2.27) with $C = 4\pi r_{\max}^{1/2}$.

It remains to establish the estimate (3.2.23) in case $\kappa = 1$. For this we write

$$\mathrm{c}_S^\kappa(m,n) = (-4\pi^2)^\kappa n^{(2\kappa-1)}\, \mathrm{Kl}_S^\kappa(m,n) + \mathrm{c}_S^\kappa(m,n)_+ \tag{3.2.29}$$

where

$$\mathrm{c}_S^\kappa(m,n)_+ = (-4\pi^2)^\kappa \sum_{k>0} (-4\pi^2 m)^{(k)} n^{(k+2\kappa-1)}\, \mathrm{Kl}_S^{\kappa+k}(m,n) \tag{3.2.30}$$

is an absolutely convergent sum (cf. (3.2.26)) and we have

$$|\mathrm{c}_S^\kappa(m,n)_+| = O\left(|m|^{1/2-\kappa}|n|^{\kappa-1/2} e^{C|mn|^{1/2}}\right) \tag{3.2.31}$$

with $C = 4\pi r_{\max}^{1/2}$ according to the estimates (3.2.27) and (3.2.28). Now we see from (3.2.29) that the estimate (3.2.31) also holds for $\mathrm{c}_S^\kappa(m,n)$ since (3.2.6) (or (3.2.11) in case $m = 0$) implies that $\mathrm{Kl}_S^\kappa(m,n)$ has at most polynomial growth in n as $|n| \to \infty$ when $\kappa = 1$. This completes the proof. □

We write $\mathrm{c}_{\Gamma,\mathsf{p}|\mathsf{q}}^\kappa(m,n)$ for $\mathrm{c}_S^\kappa(m,n)$ in the case that $S = [\![\Sigma_\mathsf{p}^{-1}\Sigma_\mathsf{q}]\!]^\times$. This notation suppresses the dependence on the choice of scaling cosets; the following lemma describes this dependence explicitly.

Lemma 3.2.9. *Let $\mathsf{p},\mathsf{q} \in \mathsf{P}_\Gamma$, let Σ_p and Σ'_p be scaling cosets for Γ at p and let Σ_q and Σ'_q be scaling cosets for Γ at q. Set $S = [\![\Sigma_\mathsf{p}^{-1}\Sigma_\mathsf{q}]\!]^\times$ and $S' = [\![(\Sigma'_\mathsf{p})^{-1}\Sigma'_\mathsf{q}]\!]^\times$. Then we have $\Sigma'_\mathsf{p} = \Sigma_\mathsf{p} T^\alpha$*

and $\Sigma'_{\mathsf{q}} = \Sigma_{\mathsf{q}} T^{\beta}$ *and* $\mathrm{c}^{\kappa}_{S'}(m,n) = \mathrm{e}(m\alpha + n\beta)\mathrm{c}^{\kappa}_{S}(m,n)$ *for some* $\alpha, \beta \in \mathbb{Q}$.

With a fixed choice of subset $S \subset [\![G(\mathbb{Q})]\!]^{\times}$ and integer $\kappa \in \mathbb{Z}$, we assemble the coefficient functions $\mathrm{c}^{\kappa}_{S}(m,n)$ into a formal series $\tilde{F}^{\kappa}_{S}(p,q)$ by setting

$$\tilde{F}^{\kappa}_{S}(p,q) = \sum_{m,n\in\mathbb{Z}} \mathrm{c}^{\kappa}_{S}(m,n) p^m q^n. \tag{3.2.32}$$

The formal Fourier coefficients of the formal series $\tilde{F}^{\kappa}_{S}(p,q)$ with respect to the variable p are of particular importance. For fixed $S \subset [\![G(\mathbb{Q})]\!]^{\times}$ and $\kappa \in \mathbb{Z}$ we define formal series $\tilde{F}^{\kappa(m)}_{S}(q)$ by requiring that

$$\tilde{F}^{\kappa}_{S}(p,q) = \sum_{m\in\mathbb{Z}} \tilde{F}^{\kappa(m)}_{S}(q) p^m. \tag{3.2.33}$$

We define the *singular* and *regular part of* $\tilde{F}^{\kappa(m)}_{S}(q)$, to be denoted $\tilde{F}^{\kappa(m)}_{S}(q)_{\mathrm{sing}}$ and $\tilde{F}^{\kappa(m)}_{S}(q)_{\mathrm{reg}}$, respectively, by setting

$$\tilde{F}^{\kappa(m)}_{S}(q)_{\mathrm{sing}} = \sum_{n\geq 0} \mathrm{c}^{\kappa}_{S}(m,-n-1) q^{-n-1}, \quad \tilde{F}^{\kappa(m)}_{S}(q)_{\mathrm{reg}} = \sum_{n\geq 0} \mathrm{c}^{\kappa}_{S}(m,n) q^{n}. \tag{3.2.34}$$

We define $\tilde{F}^{\kappa(m)}_{S}(q)_{\mathrm{van}}$ to be the series obtained by removing the constant term from $\tilde{F}^{\kappa(m)}_{S}(q)_{\mathrm{reg}}$.

$$\tilde{F}^{\kappa(m)}_{S}(q)_{\mathrm{van}} = \sum_{n>0} \mathrm{c}^{\kappa}_{S}(m,n) q^{n} \tag{3.2.35}$$

Given the formal series $\tilde{F}^{\kappa}_{S}(p,q)$ for some S and κ, one may ask if the coefficient series $\tilde{F}^{\kappa(m)}_{S}(q)$ define holomorphic functions on $\mathbb{H}$ upon the substitution of $\mathrm{e}(z)$ for q. Supposing this to be the case, we define $F^{\kappa(m)}_{S}(z)$ by setting

$$F^{\kappa(m)}_{S}(z) = \tilde{F}^{\kappa(m)}_{S}(\mathrm{e}(z)) = \sum_{n\in\mathbb{Z}} \mathrm{c}^{\kappa}_{S}(m,n)\, \mathrm{e}(nz), \tag{3.2.36}$$

and we define $F^{\kappa(m)}_{S}(z)_{\mathrm{sing}}$, $F^{\kappa(m)}_{S}(z)_{\mathrm{reg}}$ and $F^{\kappa(m)}_{S}(z)_{\mathrm{van}}$ in the analogous way.

We write $F^{\kappa(m)}_{\Gamma,\mathsf{p}|\mathsf{q}}(z)$ for $F^{\kappa(m)}_{S}(z)$ in the case that $S = \left[\!\left[\Sigma_{\mathsf{p}}^{-1}\Sigma_{\mathsf{q}}\right]\!\right]^{\times}$ for Γ a group commensurable with $G(\mathbb{Z})$ and for $\Sigma_{\mathsf{p}}, \Sigma_{\mathsf{q}} \in \Gamma\backslash G(\mathbb{Q})$ scaling cosets for Γ at cusps $\mathsf{p}, \mathsf{q} \in \mathsf{P}_{\Gamma}$. We write $F^{(m)}_{S}(z)$ as a shorthand for $F^{0(m)}_{S}(z)$, and interpret the notation $F^{(m)}_{\Gamma,\mathsf{p}|\mathsf{q}}(z)$, &c., similarly.

The regular part of the formal series $\tilde{F}^{\kappa(m)}_{S}(q)$ indeed defines a holomorphic function on $\mathbb{H}$, for suitable $\kappa, m \in \mathbb{Z}$, in the case that S is of the form $\left[\!\left[\Sigma_{\mathsf{p}}^{-1}\Sigma_{\mathsf{q}}\right]\!\right]^{\times}$.

Proposition 3.2.10. *Let* Γ *be a group commensurable with* $G(\mathbb{Z})$*, let* $\mathsf{p}, \mathsf{q} \in \mathsf{P}_{\Gamma}$ *be cusps of* Γ

and let $\kappa, m \in \mathbb{Z}$. *Then the series*

$$F^{\kappa(m)}_{\Gamma,\mathsf{p}|\mathsf{q}}(z)_{\mathrm{van}} = \sum_{n>0} \mathrm{c}^{\kappa}_{\Gamma,\mathsf{p}|\mathsf{q}}(m,n)\,\mathrm{e}(nz) \tag{3.2.37}$$

converges absolutely and locally uniformly for $z \in \mathbb{H}$. *In particular,* $F^{\kappa(m)}_{\Gamma,\mathsf{p}|\mathsf{q}}(z)_{\mathrm{van}}$ *is a holomorphic function on* $\mathbb{H}$ *for any* $\kappa, m \in \mathbb{Z}$.

Proof. For $m \neq 0$ the claim follows directly from the estimate (3.2.23). If $m = 0$ and $\kappa \leq 0$ then $F^{\kappa(m)}_{\Gamma,\mathsf{p}|\mathsf{q}}(z)_{\mathrm{van}}$ vanishes identically according to (3.2.21). For $m = 0$ and $\kappa > 0$ the required convergence of (3.2.37) follows from (3.2.24) since $\mathrm{Kl}^{\kappa}_{S}(0,n)$ has at most polynomial growth in n according to the proof of Proposition 3.2.1. □

By a similar argument to that just given we see that the singular part of the formal series $\tilde{F}^{\kappa(m)}_{S}(q)$ defines an anti-holomorphic function on $\mathbb{H}$ when S is of the form $\left[\!\left[\Sigma_{\mathsf{p}}^{-1}\Sigma_{\mathsf{q}}\right]\!\right]^{\times}$.

Proposition 3.2.11. *Let* Γ *be a group commensurable with* $G(\mathbb{Z})$, *let* $\mathsf{p},\mathsf{q} \in \mathsf{P}_{\Gamma}$ *be cusps of* Γ *and let* $\kappa, m \in \mathbb{Z}$. *Then the series*

$$F^{\kappa(m)}_{\Gamma,\mathsf{p}|\mathsf{q}}(\bar{z})_{\mathrm{sing}} = \sum_{n>0} \mathrm{c}^{\kappa}_{\Gamma,\mathsf{p}|\mathsf{q}}(m,-n)\,\mathrm{e}(-n\bar{z}) \tag{3.2.38}$$

converges absolutely and locally uniformly for $z \in \mathbb{H}$. *In particular,* $F^{\kappa(m)}_{\Gamma,\mathsf{p}|\mathsf{q}}(\bar{z})_{\mathrm{sing}}$ *is an anti-holomorphic function on* $\mathbb{H}$ *for any* $\kappa, m \in \mathbb{Z}$.

The notation $F^{\kappa(m)}_{\Gamma,\mathsf{p}|\mathsf{q}}$ suppresses the dependence upon the choice of scaling cosets Σ_{p} and Σ_{q}. The next result encodes this dependence precisely.

Proposition 3.2.12. *Let* Γ *be a group commensurable with* $G(\mathbb{Z})$, *and let* $\mathsf{p},\mathsf{q} \in \mathsf{P}_{\Gamma}$. *Let* Σ_{p} *and* Σ'_{p} *be scaling cosets for* Γ *at* p *and let* Σ_{q} *and* Σ'_{q} *be scaling cosets for* Γ *at* q. *Set* $S = \left[\!\left[\Sigma_{\mathsf{p}}^{-1}\Sigma_{\mathsf{q}}\right]\!\right]^{\times}$ *and* $S' = \left[\!\left[(\Sigma'_{\mathsf{p}})^{-1}\Sigma'_{\mathsf{q}}\right]\!\right]^{\times}$. *Then we have* $\Sigma'_{\mathsf{p}} = \Sigma_{\mathsf{p}}T^{\alpha}$ *and* $\Sigma'_{\mathsf{q}} = \Sigma_{\mathsf{q}}T^{\beta}$ *and* $F^{\kappa(m)}_{S'}(z) = \mathrm{e}(m\alpha)F^{\kappa(m)}_{S}(z+\beta)$ *for some* $\alpha, \beta \in \mathbb{Q}$.

3.3 Convergence

Our main goal in this section is to establish the convergence, and Fourier series expansion, of the Rademacher sum $R^{\kappa(m)}_{\Gamma,\mathsf{p}|\mathsf{q}}(z)$ (cf. §3.1), attached to a group Γ commensurable with $G(\mathbb{Z})$, and cusps p and q for Γ. We will require mostly minor modifications of the arguments furnished by Niebur in [Nie74]. An important technical tool in these arguments is the identity

$$\sum_{n>0} n^{(s-1)}\,\mathrm{e}(nz) = \sum_{n\in\mathbb{Z}} (-2\pi\mathbf{i})^{-s}(z+n)^{-s}, \tag{3.3.1}$$

valid for $\Re(s) > 1$, in which both sums converge absolutely and uniformly in z on compact subsets of $\mathbb{H}$. A nice proof of (3.3.1) appears in [KR01]. This identity can be extended to $s = 1$ at the expense of absolute convergence.

Lemma 3.3.1. *Let $z \in \mathbb{H}$ and let $K \in \mathbb{Z}^+$ such that $K + 1/2 > \Re(z)$. Then we have*

$$\sum_{n>0} \mathrm{e}(nz) = \sum_{\substack{n\in\mathbb{Z} \\ -K\le n\le K}} (-2\pi\mathbf{i})^{-1}(z+n)^{-1} + \lambda_K(z) \tag{3.3.2}$$

where the error term $\lambda_K(z)$ is given by

$$\lambda_K(z) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \left(\frac{1}{z+\mathbf{i}v-K-1/2} - \frac{1}{z+\mathbf{i}v+K+1/2}\right)\frac{1}{1+e^{-2\pi v}}\mathrm{d}v. \tag{3.3.3}$$

Proof. The proof follows that of (the more general) Lemma 4.1 in [Nie74]. Define a function $f(\xi)$ by setting $f(\xi) = (2\pi\mathbf{i}\xi)^{-1}(\mathrm{e}(\xi - z) - 1)^{-1}$. Then $f(\xi)$ has a pole at $\xi = 0$ and a pole also at $\xi = z + n$ for each $n \in \mathbb{Z}$. For the residue of $f(\xi)\mathrm{d}\xi$ at $\xi = 0$ we have

$$\mathrm{Res}_{\xi=0} f(\xi)\mathrm{d}\xi = \frac{1}{2\pi\mathbf{i}}\frac{1}{\mathrm{e}(-z)-1} = \frac{1}{2\pi\mathbf{i}}\sum_{n>0}\mathrm{e}(nz), \tag{3.3.4}$$

which is $1/2\pi\mathbf{i}$ times the left hand side of (3.3.2), while the residue at $\xi = z+n$ is $(2\pi\mathbf{i})^{-2}(z+n)^{-1}$. Let K be a positive integer such that $K + 1/2 > \Re(z)$ and set $Q = K + 1/2$ for convenience. Let γ be the positively oriented boundary of the rectangle with corners $\Re(z) \pm Q \pm \mathbf{i}L$ where $L > \Im(z)$. Then by the residue theorem we have

$$\int_\gamma f(\xi)\mathrm{d}\xi = \sum_{n>0}\mathrm{e}(nz) + \sum_{-K\le n\le K}(2\pi\mathbf{i})^{-1}(z+n)^{-1}. \tag{3.3.5}$$

On the other hand, the contributions of the horizontal edges of γ to $\int_\gamma f(\xi)\mathrm{d}\xi$ tend to 0 as $L \to \infty$ since $f(\xi)$ decays exponentially as $\Im(\xi) \to \infty$. The residues are independent of L for L sufficiently large so taking the limit as $L \to \infty$ we obtain

$$\sum_{n>0}\mathrm{e}(nz) + \sum_{-K\le n\le K}(2\pi\mathbf{i})^{-1}(z+n)^{-1} = \lambda_K(z) \tag{3.3.6}$$

where $\lambda_K(z)$ is the limit as $L \to \infty$ of the contributions of the vertical edges of γ. Setting $x = \Re(z)$ we have

$$\lambda_K(z) = \mathbf{i}\int_{-\infty}^{\infty}(f(x+Q+\mathbf{i}v) - f(x-Q+\mathbf{i}v))\mathrm{d}v. \tag{3.3.7}$$

After replacing v with $v + \Im(z)$ and observing that $\mathrm{e}(Q) = \mathrm{e}(-Q) = -1$ since K is an integer

we obtain the expression

$$\lambda_K(z) = \frac{1}{2\pi}\int_{-\infty}^{\infty}\left(\frac{1}{z+\mathbf{i}v-Q}-\frac{1}{z+\mathbf{i}v+Q}\right)\frac{1}{1+\mathrm{e}(\mathbf{i}v)}\mathrm{d}v \tag{3.3.8}$$

and combining this with (3.3.6) recovers the required identity. □

Setting $Q = K + 1/2$ and substituting Qv for v in (3.3.3) we find that

$$\begin{aligned}\lambda_K(z) &= \frac{1}{2\pi}\int_{-\infty}^{\infty}\left(\frac{1}{z/Q+\mathbf{i}v-1}-\frac{1}{z/Q+\mathbf{i}v+1}\right)\frac{1}{1+\mathrm{e}(\mathbf{i}Qv)}\mathrm{d}v\\ &= \frac{1}{\pi}\int_{-\infty}^{\infty}\frac{1}{((z/Q+\mathbf{i}v)^2-1)}\frac{1}{1+e^{-2\pi Qv}}\mathrm{d}v\end{aligned} \tag{3.3.9}$$

which demonstrates that $\lambda_K(z)$ tends to $-1/\pi$ times $\int_0^\infty(1+v^2)^{-1}\mathrm{d}v = \pi/2$ as $K\to\infty$. We thus obtain

$$\sum_{n>0}\mathrm{e}(nz) = -\frac{1}{2}+\lim_{K\to\infty}\sum_{-K<n<K}(-2\pi\mathbf{i})^{-1}(z+n)^{-1}. \tag{3.3.10}$$

The formulas (3.3.1) and (3.3.10) are collectively known as the *Lipschitz summation formula*.

Theorem 3.3.2. *Let Γ be a group commensurable with $G(\mathbb{Z})$, and let $\mathsf{p},\mathsf{q}\in\mathsf{P}_\Gamma$ be cusps of Γ. Let $\kappa, m\in\mathbb{Z}$ such that $\kappa\le 0$ and $m<0$. Then the limit defining the classical Rademacher sum $R^{\kappa(m)}_{\Gamma,\mathsf{p}|\mathsf{q}}(z)$ converges locally uniformly to a holomorphic function on $\mathbb{H}$, and we have*

$$R^{\kappa(m)}_{\Gamma,\mathsf{p}|\mathsf{q}}(z) = \delta_{\Gamma,\mathsf{p}|\mathsf{q}}\,\mathrm{e}(mz)+\frac{1}{2}\mathrm{c}^{\kappa}_{\Gamma,\mathsf{p}|\mathsf{q}}(m,0)+F^{\kappa(m)}_{\Gamma,\mathsf{p}|\mathsf{q}}(z)_{\mathrm{van}}. \tag{3.3.11}$$

Proof. Set $U = \left[\!\left[\Sigma_{\mathsf{p}}^{-1}\Sigma_{\mathsf{q}}\right]\!\right]$ and $S = \left[\!\left[\Sigma_{\mathsf{p}}^{-1}\Sigma_{\mathsf{q}}\right]\!\right]^{\times}$. For $\kappa<0$ the result follows from Lemmas 4.2 and 4.3 of [Nie74]. Indeed, using Lemma 3.1.1 we may identify the Rademacher component function $R^{\kappa(m)}_{[\![\chi]\!]}(z)$ with the function $s_{[\![\chi]\!]}(z)+t_{[\![\chi]\!]}(z)$ defined on page 376 of [Nie74]. Then the expression

$$\sum_{c\in\mathbb{Z}^+}\lim_{K\to\infty}\sum_{\substack{[\![\chi]\!]\in U_{\le K}\\ c(\chi)=c}}R^{\kappa(m)}_{[\![\chi]\!]}(z) \tag{3.3.12}$$

is, up to addition by a certain constant function, the definition of Rademacher sum used by Niebur in loc. cit. The Lipschitz summation formula(s) (3.3.1) and (3.3.10) are the main tools used in equating (3.3.12) with the right hand side of (3.3.11).

In case $\kappa = 0$ we require to reinforce the argument of [Nie74] with a non-trivial estimate for the Selberg–Kloosterman zeta function $\mathrm{Kl}^s_S(m,n)$ at $s=1$; we will apply the result (3.2.6) of

Goldfeld–Sarnark to this purpose. Assume then that $\kappa = 0$ and observe that we have

$$\begin{aligned} R^{\kappa(m)}_{\Gamma,\mathfrak{p}|\mathfrak{q}}(z) &= \delta_{\Gamma,\mathfrak{p}|\mathfrak{q}}\,\mathrm{e}(mz) + \lim_{K\to\infty} \sum_{[\![\chi\in U^\times_{\le K}} \mathrm{e}(m\,[\![\chi\cdot\infty)\,\mathrm{e}(m\,[\![\chi\cdot z - m\,[\![\chi\cdot\infty, 1) \\ &= \delta_{\Gamma,\mathfrak{p}|\mathfrak{q}}\,\mathrm{e}(mz) + \lim_{K\to\infty} \sum_{[\![\chi\in U^\times_{\le K}} \mathrm{e}(m\,[\![\chi\cdot\infty) \sum_{k\ge 0} (-2\pi\mathbf{i}\varrho\,[\![\chi]\!])^{k+1} m^{(k+1)} (z - \chi^{-1}]\!]\cdot\infty)^{-k-1} \end{aligned} \tag{3.3.13}$$

according to (3.1.7), the definition (2.1.6) of the generalized exponential $\mathrm{e}(z,s)$, and the identity (2.4.4). We write $R^{\kappa(m)}_{\Gamma,\mathfrak{p}|\mathfrak{q}}(z) = \mathrm{e}(mz) + R_0 + R_+$ where

$$R_0 = \lim_{K\to\infty} \sum_{[\![\chi\in U^\times_{\le K}} \mathrm{e}(m\,[\![\chi\cdot\infty)(-2\pi\mathbf{i}\varrho\,[\![\chi]\!])m(z-\chi^{-1}]\!]\cdot\infty)^{-1}, \tag{3.3.14}$$

$$R_+ = \sum_{[\![\chi\in U^\times} \mathrm{e}(m\,[\![\chi\cdot\infty) \sum_{k>0} (-2\pi\mathbf{i}\varrho\,[\![\chi]\!])^{k+1} m^{(k+1)} (z-\chi^{-1}]\!]\cdot\infty)^{-k-1}, \tag{3.3.15}$$

and the second sum R_+ is absolutely convergent, locally uniformly for $z\in\mathbb{H}$. After choosing a representative χ for each double coset $[\![\chi]\!]$ in S we may rewrite R_+ as

$$R_+ = \sum_{[\![\chi]\!]\in S} \mathrm{e}(m\,[\![\chi]\!]\cdot\infty) \sum_{k>0} (-2\pi\mathbf{i}\varrho\,[\![\chi]\!])^{k+1} m^{(k+1)} \sum_{n\in\mathbb{Z}} (z-\chi^{-1}]\!]\cdot\infty + n)^{-k-1}, \tag{3.3.16}$$

and applying the Lipschitz summation formula (3.3.1) to this we obtain

$$\begin{aligned} R_+ &= \sum_{n>0} \sum_{[\![\chi]\!]\in S} \mathrm{e}(m\,[\![\chi]\!]\cdot\infty)\,\mathrm{e}(-n\,[\![\chi^{-1}]\!]\cdot\infty) \sum_{k>0} (-4\pi^2\varrho\,[\![\chi]\!])^{k+1} m^{(k+1)} n^{(k)}\,\mathrm{e}(nz) \\ &= \sum_{n>0} \mathrm{c}^\kappa_S(m,n)_+\,\mathrm{e}(nz) \end{aligned} \tag{3.3.17}$$

where $\mathrm{c}^\kappa_S(m,n)_+$ satisfies $\mathrm{c}^\kappa_S(m,n) = (-4\pi^2)m\,\mathrm{Kl}^1_S(m,n) + \mathrm{c}^\kappa_S(m,n)_+$ according to (3.2.21). For K a positive integer set

$$R_0(K) = (-2\pi\mathbf{i})m \sum_{U^\times_{\le K}} \mathrm{e}(m\,[\![\chi\cdot\infty)\varrho\,[\![\chi]\!]\,(z-\chi^{-1}]\!]\cdot\infty)^{-1} \tag{3.3.18}$$

so that $R_0 = \lim_{K\to\infty} R_0(K)$. Then choosing as above a representative χ for each $[\![\chi]\!]\in S$ we

have

$$R_0(K) = (-2\pi\mathbf{i})m \sum_{[\![\chi]\!]\in S_{\leq K}} \mathrm{e}(m\,[\![\chi]\!]\cdot\infty)\varrho\,[\![\chi]\!] \sum_{\substack{n\in\mathbb{Z}\\ -K^2\leq d(\chi)+nc(\chi)\leq K^2}} (z-\chi^{-1}[\![\chi]\!]\cdot\infty+n)^{-1}$$
$$= (-2\pi\mathbf{i})m \sum_{0<c<K}\sum_{[\![\chi]\!]\in S_c} \mathrm{e}(m\,[\![\chi]\!]\cdot\infty)\varrho\,[\![\chi]\!] \left(\sum_{|n|\leq K^2/c} (z-\chi^{-1}[\![\chi]\!]\cdot\infty+n)^{-1} + O(c/K^2)\right) \tag{3.3.19}$$

where the term $O(c/K^2)$ accounts for the discrepency between the two summations over n. Applying the Lipschitz summation formula (3.3.6) to the second line of (3.3.19) we obtain

$$R_0(K) = (-4\pi^2)m \sum_{0<c<K}\sum_{[\![\chi]\!]\in S_c} \mathrm{e}(m\,[\![\chi]\!]\cdot\infty)\varrho\,[\![\chi]\!] \left(\frac{1}{2} + \sum_{n>0}\mathrm{e}(-n\,[\![\chi^{-1}]\!]\cdot\infty)\,\mathrm{e}(nz) + O(c/K^2)\right) \tag{3.3.20}$$

since the error term $\lambda_X(z)$ appearing in (3.3.6) satisfies $\lambda_X(z)+1/2 = O(1/X)$. We conclude from this that

$$R_0 = \lim_{K\to\infty} \sum_{[\![\chi]\!]\in S_{\leq K}} \mathrm{e}(m\,[\![\chi]\!]\cdot\infty)(-4\pi^2)\varrho\,[\![\chi]\!]\, m \left(\frac{1}{2} + \sum_{n>0}\mathrm{e}(-n\,[\![\chi^{-1}]\!]\cdot\infty)\,\mathrm{e}(nz)\right). \tag{3.3.21}$$

We claim now that $R_0 = \mathrm{c}_S^\kappa(m,0)/2 + \sum_{n>0}\mathrm{c}_S^\kappa(m,n)_0\,\mathrm{e}(nz)$ where $\mathrm{c}_S^\kappa(m,n)_0 = \mathrm{c}_S^\kappa(m,n) - \mathrm{c}_S^\kappa(m,n)_+ = (-4\pi^2)m\,\mathrm{Kl}_S^1(m,n)$. To see this observe that for fixed positive K we have

$$\sum_{[\![\chi]\!]\in S_{\leq K}} \mathrm{e}(m\,[\![\chi]\!]\cdot\infty)\varrho\,[\![\chi]\!] \sum_{n>0}\mathrm{e}(-n\,[\![\chi^{-1}]\!]\cdot\infty)\,\mathrm{e}(nz) = \sum_{n>0}\mathrm{Kl}^1_{S_{\leq K}}(m,n)\,\mathrm{e}(nz) \tag{3.3.22}$$

since the left hand side converges absolutely. Now $\mathrm{Kl}^1_{S_{\leq K}}(m,n) = \mathrm{Kl}^1_S(m,n) + O(n)$ as $K\to\infty$ according to (3.2.10) since the $\Sigma(x)$ there is just $\mathrm{Kl}^1_{S_{\leq x}}(m,n)$ (cf. (3.2.9)). Then (3.2.6) implies $\mathrm{Kl}^1_{S_{\leq K}}(m,n) = O(n)$ uniformly in K and n and we conclude that the limit as $K\to\infty$ of the right hand side of (3.3.22) is $\sum_{n>0}\mathrm{Kl}^1_S(m,n)\,\mathrm{e}(nz)$, and thus

$$R_0 = \mathrm{c}_S^\kappa(m,0)/2 + \sum_{n>0}\mathrm{c}_S^\kappa(m,n)_0\,\mathrm{e}(nz) \tag{3.3.23}$$

as was claimed. Taking (3.3.23) together with (3.3.17) we arrive at the required identity (3.3.11) in case $\kappa = 0$. The proof is complete. □

The next result identifies the Fourier series expansion of the conjugate Rademacher sum $\bar{R}^{\kappa(m)}_{\Gamma,\mathfrak{p}|\mathfrak{q}}(z)$. As such it is a natural counterpart to Theorem 3.3.2. The method of proof is directly

analogous to that of Theorem 3.3.2.

Theorem 3.3.3. *Let Γ be a group commensurable with $G(\mathbb{Z})$, and let $\mathsf{p},\mathsf{q} \in \mathsf{P}_\Gamma$ be cusps of Γ. Let $\kappa, m \in \mathbb{Z}$ such that $\kappa \leq 0$ and $m < 0$. Then the limit defining the conjugate Rademacher sum $\bar{R}^{\kappa(m)}_{\Gamma,\mathsf{p}|\mathsf{q}}(z)$ converges locally uniformly to an anti-holomorphic function on $\mathbb{H}$, and we have*

$$\bar{R}^{\kappa(m)}_{\Gamma,\mathsf{p}|\mathsf{q}}(z) = \delta_{\Gamma,\mathsf{p}|\mathsf{q}}\, \mathrm{e}(m\bar{z}) - \frac{1}{2}\mathrm{c}^{\kappa}_{\Gamma,\mathsf{p}|\mathsf{q}}(m,0) - F^{\kappa(m)}_{\Gamma,\mathsf{p}|\mathsf{q}}(\bar{z})_{\mathrm{sing}}. \tag{3.3.24}$$

Theorems 3.3.2 and 3.3.3 furnish the Fourier expansions of the classical and conjugate Rademacher sums of non-positive weight. Directly analogous methods can be used to determine expressions for the Fourier coefficients of the Rademacher sums of positive weight, which are, after all, just the holomorphic Poincaré series (cf. §3.1). Indeed, there are fewer technical difficulties in the case of positive weight, since the sum appearing in the definition (3.1.2) of the Poincaré series $P^{\kappa(m)}_{\Gamma,\mathsf{p}|\mathsf{q}}$ is absolutely and locally uniformly convergent (at least when $\kappa > 1$). For completeness we state an analogue of Theorem 3.3.2 for $\kappa > 0$. The result is standard (cf. [Iwa02]).

Theorem 3.3.4. *Let Γ be a group commensurable with $G(\mathbb{Z})$, and let $\mathsf{p},\mathsf{q} \in \mathsf{P}_\Gamma$ be cusps of Γ. Let $\kappa, m \in \mathbb{Z}$ such that $\kappa > 0$ and $m \geq 0$. Then for $z \in \mathbb{H}$ we have*

$$R^{\kappa(m)}_{\Gamma,\mathsf{p}|\mathsf{q}}(z) = \delta_{\Gamma,\mathsf{p}|\mathsf{q}}\, \mathrm{e}(mz) + F^{\kappa(m)}_{\Gamma,\mathsf{p}|\mathsf{q}}(z)_{\mathrm{van}}. \tag{3.3.25}$$

Recall from §3.1 that we write $R^{(m)}_{\Gamma,\mathsf{p}|\mathsf{q}}(z)$ for $R^{0(m)}_{\Gamma,\mathsf{p}|\mathsf{q}}(z)$ and interpret $\bar{R}^{(m)}_{\Gamma,\mathsf{p}|\mathsf{q}}(z)$ similarly. To conclude this section we consider the difference $R^{(m)}_{\Gamma,\mathsf{p}|\mathsf{q}}(z) - \bar{R}^{(m)}_{\Gamma,\mathsf{p}|\mathsf{q}}(z)$. Combining Theorems 3.3.2 and 3.3.3 we obtain the following identification of the Fourier expansion of this function in terms of the exponential function $\mathrm{e}(z)$ and its conjugate $\mathrm{e}(-\bar{z})$.

Theorem 3.3.5. *Let Γ be a group commensurable with $G(\mathbb{Z})$, let $\mathsf{p},\mathsf{q} \in \mathsf{P}_\Gamma$ be cusps of Γ, and let $m < 0$. Then we have*

$$\begin{aligned} R^{(m)}_{\Gamma,\mathsf{p}|\mathsf{q}}(z) - \bar{R}^{(m)}_{\Gamma,\mathsf{p}|\mathsf{q}}(z) = \delta_{\Gamma,\mathsf{p}|\mathsf{q}}(\mathrm{e}(mz) - \mathrm{e}(m\bar{z})) \\ + F^{(m)}_{\Gamma,\mathsf{p}|\mathsf{q}}(\bar{z})_{\mathrm{sing}} + \mathrm{c}_{\Gamma,\mathsf{p}|\mathsf{q}}(m,0) + F^{(m)}_{\Gamma,\mathsf{p}|\mathsf{q}}(z)_{\mathrm{van}} \end{aligned} \tag{3.3.26}$$

for $z \in \mathbb{H}$.

3.4 Variance

Recall from §3.1 that we write $R^{\kappa(m)}_{\Gamma,\mathsf{p}}(z)$ for $R^{\kappa(m)}_{\Gamma,\mathsf{p}|\mathsf{q}}(z)$ in case q is the infinite cusp $\Gamma\cdot\infty$. Suppose that Γ is commensurable with the modular group and has width one at infinity (cf. §2.6). We will establish in this section that the Rademacher sum $R^{\kappa(m)}_{\Gamma,\mathsf{p}}(z)$, once corrected by the addition of a certain constant function, is an automorphic integral of weight 2κ for Γ whenever $\kappa = 1$

and $m > 0$ or $\kappa > 1$ and $m \geq 0$ or $\kappa \leq 0$ and $m < 0$. This is quite general since, according to the discussion of §3.1, any Rademacher sum $R^{\kappa(m)}_{\Gamma,\mathsf{p}|\mathsf{q}}(z)$ can be expressed in the form $R^{\kappa(m)}_{\Gamma',\mathsf{p}'}(z)$ for some group Γ' having width one at infinity (cf. (3.1.12)).

Let us begin by considering the weight 2κ action of $G(\mathbb{Q})$ on the Rademacher component function $R^{\kappa(m)}_{[\![\chi}(z)$ of (3.1.4).

Lemma 3.4.1. *Let $[\![\chi \in [\![G(\mathbb{Q})^{\times}$ and $\sigma \in G(\mathbb{Q})$, and let $\kappa, m \in \mathbb{Z}$. In case $\kappa > 0$ and $m \geq 0$ we have*

$$\left(R^{\kappa(m)}_{[\![\chi}\Big|^{\kappa}\sigma\right)(z) = R^{\kappa(m)}_{[\![\chi\sigma}(z), \tag{3.4.1}$$

and in case $\kappa \leq 0$ and $m < 0$ we have

$$\left(R^{\kappa(m)}_{[\![\chi}\Big|^{\kappa}\sigma\right)(z) = R^{\kappa(m)}_{[\![\chi\sigma}(z) + m^{1-2\kappa}\left(J^{\kappa}_{\infty\cdot[\![\sigma}P^{1-\kappa(-m)}_{[\![\chi\sigma}\right)(z). \tag{3.4.2}$$

Proof. Note that we have

$$\left(P^{\kappa(m)}_{[\![\chi}|^{\kappa}\sigma\right)(z) = P^{\kappa(m)}_{[\![\chi\sigma}(z) \tag{3.4.3}$$

for all $\kappa, m \in \mathbb{Z}$, where $P^{\kappa(m)}_{[\![\chi}(z) = \mathrm{e}(m\,[\![\chi\cdot z)\,\mathrm{jac}([\![\chi\,,z)^{\kappa}$ (cf. (3.1.1)), so the identity (3.4.1) follows from (3.4.3) and the fact that $R^{\kappa(m)}_{[\![\chi}(z) = P^{\kappa(m)}_{[\![\chi}(z)$ in case $\kappa > 0$, by the definition of $R^{\kappa(m)}_{[\![\chi}(z)$ (cf. §3.1). For (3.4.2) we suppose $\kappa \leq 0$ and $m < 0$ and use Lemma 3.1.1 to write the Rademacher component function $R^{\kappa(m)}_{[\![\chi}(z)$ in terms of the functions $P^{\kappa(m)}_{[\![\chi}(z)$ and the integral operator J^{κ}_{w}. Then, employing (3.4.3) and the identity (2.7.6), we compute

$$\begin{aligned}\left(R^{\kappa(m)}_{[\![\chi}\Big|^{\kappa}\sigma\right)(z) &= \left(P^{\kappa(m)}_{[\![\chi}\Big|^{\kappa}\sigma\right)(z) - m^{1-2\kappa}\left(\left(J^{\kappa}_{\infty\cdot[\![\chi}P^{1-\kappa(-m)}_{[\![\chi}\right)\Big|^{\kappa}\sigma\right)(z)\\ &= P^{\kappa(m)}_{[\![\chi\sigma}(z) - m^{1-2\kappa}\left(J^{\kappa}_{\infty\cdot[\![\chi\sigma}P^{1-\kappa(-m)}_{[\![\chi\sigma}\right)(z) + m^{1-2\kappa}\left(J^{\kappa}_{\infty\cdot[\![\sigma}P^{1-\kappa(-m)}_{[\![\chi\sigma}\right)(z).\end{aligned} \tag{3.4.4}$$

Applying Lemma 3.1.1 once again we recognize the first two terms on the right hand side (of the second line) of (3.4.4) to be $R^{\kappa(m)}_{[\![\chi\sigma}(z)$. We thus obtain the required identity (3.4.2). □

As for technical tools, in addition to Lemma 3.3.1, we also use the following result which shows that there is some flexibility in the convergence of the limit (3.1.6) defining the classical Rademacher sum.

Lemma 3.4.2. *Let Γ be a group commensurable with $G(\mathbb{Z})$, let $\mathsf{p} \in \mathsf{P}_{\Gamma}$ be a cusp of Γ and let $\sigma \in G(\mathbb{Q})$. Let $\kappa, m \in \mathbb{Z}$ such that either $\kappa = 1$ and $m > 0$ or $\kappa > 1$ and $m \geq 0$ or $\kappa \leq 0$ and*

$m < 0$. Then we have

$$\lim_{K\to\infty} \sum_{[\![\chi \in (U_{\leq K})\sigma} R^{\kappa(m)}_{[\![\chi}(z) = \lim_{K\to\infty} \sum_{[\![\chi \in (U\sigma)_{\leq K}} R^{\kappa(m)}_{[\![\chi}(z) \tag{3.4.5}$$

for $U = [\![\Sigma_{\mathsf{p}}^{-1}$ and $z \in \mathbb{H}$.

Proof. For $\kappa > 1$ and $m \geq 0$ both sums in (3.4.5) are absolutely convergent, and the identity certainly holds. In case $\kappa = 1$ and $m > 0$, or $\kappa \leq 0$ and $m < 0$, the identity (3.4.5) is established via the argument of Lemma 4.3 in [Nie74]. Note that this argument fails in the case that $\kappa = 1$ and $m = 0$, and indeed, the identity (3.4.5) does not hold in general when $\kappa = 1$ and $m = 0$. □

Theorem 3.4.3. *Let Γ be a group commensurable with $G(\mathbb{Z})$ that has width one at infinity and let $\mathsf{p} \in \mathsf{P}_\Gamma$ be a cusp of Γ. Let $\kappa, m \in \mathbb{Z}$ such that $\kappa > 1$ and $m \geq 0$ or $\kappa = 1$ and $m > 0$. Then the function $R^{\kappa(m)}_{\Gamma,\mathsf{p}}(z)$ is a modular form of weight 2κ for Γ, and is a cusp form in case $m > 0$. Further, for $\mathsf{q} \in \mathsf{P}_\Gamma$ another cusp of Γ the Fourier expansion of the function $R^{\kappa(m)}_{\Gamma,\mathsf{p}|\mathsf{q}}(z)$ is the expansion of $R^{\kappa(m)}_{\Gamma,\mathsf{p}}(z)$ at the cusp q.*

Proof. Let κ and m be as in the statement of the proposition. Since Γ has width one at infinity we may take Γ to be the scaling coset for Γ at the infinite cusp. Let Σ_{p} be a scaling coset for Γ at p and set $U = [\![\Sigma_{\mathsf{p}}^{-1}$. Then in the case that $\kappa > 1$ we have

$$R^{\kappa(m)}_{\Gamma,\mathsf{p}}(z) = \sum_{[\![\chi \in U} R^{\kappa(m)}_{[\![\chi}(z) = \sum_{[\![\chi \in U} \mathrm{e}(m\,[\![\chi \cdot z)\,\mathrm{jac}([\![\chi\,, z)^{\kappa}, \tag{3.4.6}$$

with the sum(s) in (3.4.6) converging absolutely for $z \in \mathbb{H}$. Let $\sigma \in G(\mathbb{Q})$. Then we have

$$\left(R^{\kappa(m)}_{\Gamma,\mathsf{p}}\Big|^{\kappa}\sigma\right)(z) = \sum_{[\![\chi \in U} \left(R^{\kappa(m)}_{[\![\chi}\Big|^{\kappa}\sigma\right)(z) = \sum_{[\![\chi \in U} R^{\kappa(m)}_{[\![\chi\sigma}(z) = \sum_{[\![\chi \in U\sigma} R^{\kappa(m)}_{[\![\chi}(z) \tag{3.4.7}$$

by Lemma 3.4.1, with all sums absolutely convergent, locally uniformly in z. Taking $\sigma \in \Gamma$ we see that $R^{\kappa(m)}_{\Gamma,\mathsf{p}}(z)$ is an unrestricted modular form of weight 2κ for Γ since $\Sigma_{\mathsf{p}}^{-1}\sigma = \Sigma_{\mathsf{p}}^{-1}$ for $\sigma \in \Gamma$. Taking $\mathsf{q} \in \mathsf{P}_\Gamma$ and $\sigma \in \Sigma_{\mathsf{q}}$ we see that Theorem 3.3.4 gives the Fourier expansion of $R^{\kappa(m)}_{\Gamma,\mathsf{p}}(z)$ at q, in the sense of §2.7, since $\Sigma_{\mathsf{p}}^{-1}\sigma = \Sigma_{\mathsf{p}}^{-1}\Sigma_{\mathsf{q}}$ for any $\sigma \in \Sigma_{\mathsf{q}}$. We conclude that the (a priori unrestricted) modular form $R^{\kappa(m)}_{\Gamma,\mathsf{p}}(z)$ is vanishing at every cusp $\mathsf{q} \in \mathsf{P}_\Gamma$, except possibly for $\mathsf{q} = \mathsf{p}$, and vanishes even there just when $m > 0$, and thus the claim is verified for $\kappa > 1$.

Consider now the case that $\kappa = 1$. We have

$$R^{1(m)}_{\Gamma,\mathsf{p}}(z) = \lim_{K\to\infty} \sum_{[\![\chi \in U_{\leq K}} R^{1(m)}_{[\![\chi}(z) = \lim_{K\to\infty} \sum_{[\![\chi \in U_{\leq K}} \mathrm{e}(m\,[\![\chi \cdot z)\,\mathrm{jac}([\![\chi\,, z), \tag{3.4.8}$$

and it is necessary to consider (some kind of) a limit since the convergence is not absolute in this case. Let $\sigma \in G(\mathbb{Q})$ and consider the weight 2 action of σ on $R^{1(m)}_{\Gamma,\mathsf{p}}(z)$.

$$\left(R^{1(m)}_{\Gamma,\mathsf{p}}\Big|_1 \sigma\right)(z) = \lim_{K\to\infty} \sum_{[\![\chi \in U_{\leq K}} \left(R^{1(m)}_{[\![\chi}\Big|_1 \sigma\right)(z) = \lim_{K\to\infty} \sum_{[\![\chi \in U_{\leq K}} R^{1(m)}_{[\![\chi\sigma}(z). \tag{3.4.9}$$

Taking $\sigma \in \Gamma$ and applying Lemma 3.4.2 we see that $R^{1(m)}_{\Gamma,\mathsf{p}}(z)$ is a(n unrestricted) modular form for Γ. Applying Lemma 3.4.2 in the case that $\sigma \in \Sigma_{\mathsf{q}}$ for some cusp $\mathsf{q} \in \mathsf{P}_\Gamma$, we see that the Fourier expansion of the function $R^{1(m)}_{\Gamma,\mathsf{p}|\mathsf{q}}(z)$ is the expansion of $R^{1(m)}_{\Gamma,\mathsf{p}}(z)$ at q in the sense of §2.7, so Theorem 3.3.4 applies also when $\kappa = 1$, and thus $R^{1(m)}_{\Gamma,\mathsf{p}}(z)$ is a modular form of weight 2 for Γ and even a cusp form since $m > 0$. This completes the proof. □

We now present an analogue of Theorem 3.4.3 for the case that $\kappa \leq 0$.

Theorem 3.4.4. *Let Γ be a group commensurable with $G(\mathbb{Z})$ that has width one at infinity and let $\mathsf{p} \in \mathsf{P}_\Gamma$ be a cusp of Γ. Let $\kappa, m \in \mathbb{Z}$ such that $\kappa \leq 0$ and $m < 0$. Then the function $R^{\kappa(m)}_{\Gamma,\mathsf{p}}(z) + \mathrm{c}^{\kappa}_{\Gamma,\mathsf{p}}(m,0)/2$ is an automorphic integral of weight 2κ for Γ, and for $\mathsf{q} \in \mathsf{P}_\Gamma$ another cusp of Γ the Fourier expansion of the function $R^{\kappa(m)}_{\Gamma,\mathsf{p}|\mathsf{q}}(z) + \mathrm{c}^{\kappa}_{\Gamma,\mathsf{p}|\mathsf{q}}(m,0)/2$ is the expansion of $R^{\kappa(m)}_{\Gamma,\mathsf{p}}(z) + \mathrm{c}^{\kappa}_{\Gamma,\mathsf{p}}(m,0)/2$ at the cusp q.*

Proof. Let κ and m be as in the statement of the theorem. Since Γ has width one at infinity we take Γ to be the scaling coset for Γ at the infinite cusp. Let Σ_{p} be a scaling coset for Γ at p and set $U = [\![\Sigma_{\mathsf{p}}^{-1}$. Let $\sigma \in G(\mathbb{Q})$. Using Lemma 3.4.1 we compute

$$\begin{aligned}\left(R^{\kappa(m)}_{U}\Big|^{\kappa} \sigma\right)(z) &= \lim_{K\to\infty} \sum_{[\![\chi \in U_{\leq K}} R^{\kappa(m)}_{[\![\chi\sigma}(z) + m^{1-2\kappa}\left(J^{\kappa}_{\infty\cdot[\![\sigma} P^{1-\kappa(-m)}_{[\![\chi\sigma}\right)(z) \\ &= \lim_{K\to\infty} \sum_{[\![\chi \in (U_{\leq K})\sigma} R^{\kappa(m)}_{[\![\chi}(z) + m^{1-2\kappa}\left(J^{\kappa}_{\infty\cdot[\![\sigma} P^{1-\kappa(-m)}_{[\![\chi}\right)(z)\end{aligned} \tag{3.4.10}$$

for the weight 2κ action of σ on $R^{\kappa(m)}_{U}(z) = R^{\kappa(m)}_{\Gamma,\mathsf{p}}(z)$. Now Lemma 3.4.2 shows that

$$\left(R^{\kappa(m)}_{U}\Big|^{\kappa} \sigma\right)(z) = R^{\kappa(m)}_{U\sigma}(z) + m^{1-2\kappa} \lim_{K\to\infty} \sum_{[\![\chi \in (U_{\leq K})\sigma} \left(J^{\kappa}_{\infty\cdot[\![\sigma} P^{1-\kappa(-m)}_{[\![\chi}\right)(z), \tag{3.4.11}$$

suggesting that $R^{\kappa(m)}_{U}(z)$ might be an automorphic integral of weight 2κ for Γ with associated cusp form $-m^{1-2\kappa}P^{1-\kappa(-m)}_{U}(z)$ (cf. §2.7). This is not accurate, however, since the limit in (3.4.11) does not in general commute with the integral operator $J^{\kappa}_{\infty\cdot[\![\sigma}$. Rather, by the argument

of Lemma 4.4. in [Nie74], we have the identity

$$\lim_{K\to\infty}\left(J^{\kappa}_{\infty\cdot[\![\sigma}\left(\sum_{[\![\chi\in(U_{\leq K})\sigma}P^{1-\kappa(-m)}_{[\![\chi}\right)\right)(z)-\left(J^{\kappa}_{\infty\cdot[\![\sigma}\left(\lim_{K\to\infty}\sum_{[\![\chi\in(U_{\leq K})\sigma}P^{1-\kappa(-m)}_{[\![\chi}\right)\right)(z)$$
$$=-\frac{m^{2\kappa-1}}{2}\left(\mathrm{c}^{\kappa}_{S}(m,0)\,\mathrm{jac}(\sigma,z)^{\kappa}-\mathrm{c}^{\kappa}_{S'}(m,0)\right) \tag{3.4.12}$$

where $S=\{[\![\chi]\!]\mid[\![\chi\in U^{\times}\}$ and $S'=\{[\![\chi]\!]\mid[\![\chi\in(U\sigma)^{\times}\}$. Lemma 3.3.1 plays a crucial rôle in the verification of (3.4.12). Applying (3.4.12) to (3.4.11) we find that

$$\left(R^{\kappa(m)}_{U}+\frac{1}{2}\mathrm{c}^{\kappa}_{S}(m,0)\right)\Big|^{\kappa}\sigma=R^{\kappa(m)}_{U'}+\frac{1}{2}\mathrm{c}^{\kappa}_{S'}(m,0)+m^{1-2\kappa}J^{\kappa}_{\infty\cdot[\![\sigma}P^{1-\kappa(-m)}_{U'} \tag{3.4.13}$$

for $U'=U\sigma$. Taking $\sigma\in\Gamma$ and applying the identity (4.4.29) we obtain

$$\left(R^{\kappa(m)}_{\Gamma,\mathfrak{p}}+\frac{1}{2}\mathrm{c}^{\kappa}_{\Gamma,\mathfrak{p}}(m,0)\right)\Big|^{\kappa}\sigma=R^{\kappa(m)}_{\Gamma,\mathfrak{p}}+\frac{1}{2}\mathrm{c}^{\kappa}_{\Gamma,\mathfrak{p}}(m,0)+m^{1-2\kappa}J^{\kappa}_{\infty\cdot[\![\sigma}P^{1-\kappa(-m)}_{\Gamma,\mathfrak{p}} \tag{3.4.14}$$

from (3.4.13), demonstrating that the function $R^{\kappa(m)}_{\Gamma,\mathfrak{p}}(z)+\mathrm{c}^{\kappa}_{\Gamma,\mathfrak{p}}(m,0)/2$ is an unrestricted automorphic integral of weight 2κ for Γ with associated cusp form $-m^{1-2\kappa}P^{1-\kappa(-m)}_{\Gamma,\mathfrak{p}}(z)$. Letting $\mathfrak{q}\in\mathsf{P}_{\Gamma}$ and applying (3.4.13) with $\sigma\in\Sigma_{\mathfrak{q}}$ we obtain

$$\begin{aligned}R^{\kappa(m)}_{\Gamma,\mathfrak{p}|\mathfrak{q}}+\frac{1}{2}\mathrm{c}^{\kappa}_{\Gamma,\mathfrak{p}|\mathfrak{q}}(m,0)&=\left(R^{\kappa(m)}_{\Gamma,\mathfrak{p}}+\frac{1}{2}\mathrm{c}^{\kappa}_{\Gamma,\mathfrak{p}}(m,0)\right)\Big|^{\kappa}\sigma-m^{1-2\kappa}J^{\kappa}_{\infty\cdot[\![\sigma}P^{1-\kappa(-m)}_{\Gamma,\mathfrak{p}|\mathfrak{q}}\\&=\left(R^{\kappa(m)}_{\Gamma,\mathfrak{p}}+\frac{1}{2}\mathrm{c}^{\kappa}_{\Gamma,\mathfrak{p}}(m,0)+m^{1-2\kappa}J^{\kappa}_{\sigma\cdot\infty}P^{1-\kappa(-m)}_{\Gamma,\mathfrak{p}}\right)\Big|^{\kappa}\sigma,\end{aligned} \tag{3.4.15}$$

demonstrating that the Fourier expansion of the function $R^{\kappa(m)}_{\Gamma,\mathfrak{p}|\mathfrak{q}}(z)+\mathrm{c}^{\kappa}_{\Gamma,\mathfrak{p}|\mathfrak{q}}(m,0)/2$ is the expansion at $\mathfrak{q}$ of the automorphic integral $R^{\kappa(m)}_{\Gamma,\mathfrak{p}}(z)+\mathrm{c}^{\kappa}_{\Gamma,\mathfrak{p}}(m,0)/2$ in the sense of §2.7. Now Theorem 3.3.2 confirms that $R^{\kappa(m)}_{\Gamma,\mathfrak{p}}(z)+\mathrm{c}^{\kappa}_{\Gamma,\mathfrak{p}}(m,0)/2$ is meromorphic at the cusps of Γ, and is thus an automorphic integral of weight 2κ for Γ. This completes the proof of the proposition. □

We now specialize to the case that $\kappa=0$. Our final goal in this section is to establish the Γ-invariance of the harmonic function $R^{(m)}_{\Gamma,\mathfrak{p}}(z)-\bar{R}^{(m)}_{\Gamma,\mathfrak{p}}(z)$ for $m<0$ in case Γ has width one at infinity. For this we utilize the spectral theory of the hyperbolic Laplacian $\Delta=(z-\bar{z})^2\partial_z\partial_{\bar{z}}$.

Consider the *Whittaker functions* $V(z,s)$ and $W(z,s)$, defined for $z,s\in\mathbb{C}$ with $z\notin\mathbb{R}$ by

setting

$$V(z,s) = \mathrm{e}(z)\Gamma(s)\Phi(s,2s,z-\bar{z})(2\pi\mathbf{i}(\bar{z}-z))^s, \tag{3.4.16}$$

$$W(z,s) = \frac{\sin(\pi s)}{\sin(2\pi s)}\left(V(z,s) - V(z,1-s)\right), \tag{3.4.17}$$

in case $\Im(z) > 0$ where the function Φ is defined in (2.1.3). The values $V(z,s)$ for $\Im(z) < 0$ are obtained by imposing the symmetry $V(z,s) = V(\bar{z},s)$ and similarly for $W(z,s)$. Then $V(z,s)$ and $W(z,s)$ are eigenfunctions for Δ with eigenvalue $s(1-s)$, and the identity $V(z,1) = \mathrm{e}(\bar{z}) - \mathrm{e}(z)$ hints at a connection with the right hand side of (3.3.26). Define the *free space Green's function* $G(w,z,s)$, for $w,z \in \mathbb{H}$ with $w \neq z$ and $s \in \mathbb{C}$ such that $\Re(s) > 1$, by setting

$$G(w,z,s) = \frac{\Gamma(s)^2}{\Gamma(2s)}\,{}_2F_1\left(s,s;2s;h\right)h^s = \sum_{k\geq 0}\frac{\Gamma(k+s)^2}{\Gamma(k+2s)\Gamma(k+1)}h^{k+s}, \tag{3.4.18}$$

where ${}_2F_1(a,b;c;x)$ denotes the Gauss hypergeometric function and $h = h(w,z)$ is given as follows and depends only on the hyperbolic distance between w and z.

$$h(w,z) = \frac{(w-\bar{w})(z-\bar{z})}{(w-z)(\bar{w}-\bar{z})} = \frac{4\Im(w)\Im(z)}{|w-z|^2} \tag{3.4.19}$$

Then the function $z \mapsto G(w,z,s)$ satisfies $\Delta G(w,z,s) = s(1-s)G(w,z,s)$ and the integral operator with kernel $G(w,z,s)$ furnishes a right-inverse to the differential operator $f \mapsto (\Delta - \lambda)f$ when $\lambda = s(1-s)$.

Let Γ be a group commensurable with $G(\mathbb{Z})$, let p and q be cusps of Γ, and let Σ_{p} and Σ_{q} be scaling cosets for Γ at p and q, respectively. Define $G_{\Gamma,\mathsf{p}|\mathsf{q}}(w,z,s)$ for $w,z \in \mathbb{H}$ such that $\Gamma\cdot w \neq \Gamma\cdot z$ by setting

$$G_{\Gamma,\mathsf{p}|\mathsf{q}}(w,z,s) = \sum_{\chi\in\Sigma_{\mathsf{p}}^{-1}\Sigma_{\mathsf{q}}} G(w,\chi\cdot z,s). \tag{3.4.20}$$

The series (3.4.20) converges absolutely and locally uniformly when $\Re(s) > 1$ and the functions $w \mapsto G_{\Gamma,\mathsf{p}|\mathsf{q}}(w,z,s)$ and $z \mapsto G_{\Gamma,\mathsf{p}|\mathsf{q}}(w,z,s)$ are invariant for the actions of Γ^{p} and Γ^{q}, respectively. According to [Hej83] (see also [Iwa02]) we have the Fourier–Whittaker series expansion

$$\begin{aligned} &G_{\Gamma,\mathsf{p}|\mathsf{q}}(w,z,s) = \\ &\sum_{m\in\mathbb{Z}^+} G^{(m)}_{\Gamma,\mathsf{p}|\mathsf{q}}(z,s)W(-m\bar{w},s) + G^{(0)}_{\Gamma,\mathsf{p}|\mathsf{q}}(z,s)\Im(w)^{1-s} + \sum_{m\in\mathbb{Z}^+} G^{(-m)}_{\Gamma,\mathsf{p}|\mathsf{q}}(z,s)W(mw,s), \end{aligned} \tag{3.4.21}$$

converging absolutely and locally uniformly when $\Re(s) > 1$ so long as $\Im(w) > \Im(z)$ and $\Im(z)\Im(w) > \varrho_{\Gamma,\mathsf{p}|\mathsf{q}}$, for a certain constant $\varrho_{\Gamma,\mathsf{p}|\mathsf{q}}$ depending only on Γ and p and q, where

the coefficient functions $G^{(m)}_{\Gamma,\mathsf{p}|\mathsf{q}}(z,s)$ are Γ^{q}-invariant, and themselves admit the following series expansions of Poincaré type, converging absolutely for $\Re(s) > 1$ and locally uniformly for $\Im(z) > 0$.

$$G^{(m)}_{\Gamma,\mathsf{p}|\mathsf{q}}(z,s) = \frac{1}{m} \sum_{[\![\chi] \in [\![\Sigma_{\mathsf{p}}^{-1}\Sigma_{\mathsf{q}}} V(m\,[\![\chi\cdot z, s) \tag{3.4.22}$$

$$G^{(0)}_{\Gamma,\mathsf{p}|\mathsf{q}}(z,s) = \frac{4\pi}{2s-1} \sum_{[\![\chi] \in [\![\Sigma_{\mathsf{p}}^{-1}\Sigma_{\mathsf{q}}} \Im([\![\chi\cdot z)^s \tag{3.4.23}$$

$$G^{(-m)}_{\Gamma,\mathsf{p}|\mathsf{q}}(z,s) = \frac{1}{m} \sum_{[\![\chi] \in [\![\Sigma_{\mathsf{p}}^{-1}\Sigma_{\mathsf{q}}} V(-m\,[\![\chi\cdot \bar{z}, s) \tag{3.4.24}$$

Let $m > 0$. Then we have the following analogue of (3.4.21) for the Green's function coefficient $G^{(-m)}_{\Gamma,\mathsf{p}|\mathsf{q}}(z,s)$ which is again reminiscent of the right hand side of (3.3.26).

$$\begin{aligned} &G^{(-m)}_{\Gamma,\mathsf{p}|\mathsf{q}}(z,s) = \delta_{\Gamma,\mathsf{p}|\mathsf{q}}\frac{1}{m}V(-m\bar{z},s)+ \\ &\sum_{n>0} \mathrm{g}_{\Gamma,\mathsf{p}|\mathsf{q}}(-m,-n,s)W(-n\bar{z},s) + \mathrm{g}_{\Gamma,\mathsf{p}|\mathsf{q}}(-m,0,s)\Im(z)^{1-s} + \sum_{n>0} \mathrm{g}_{\Gamma,\mathsf{p}|\mathsf{q}}(-m,n,s)W(nz,s) \end{aligned} \tag{3.4.25}$$

The coefficients in (3.4.25) may be defined via the following formulas, in which we assume $m, n > 0$.

$$\begin{aligned} &\mathrm{g}_{\Gamma,\mathsf{p}|\mathsf{q}}(-m,-n,s) = \\ &\frac{1}{m^{1-s}n^s} \lim_{K\to\infty} \sum_{[\![\chi]\!] \in S_{\leq K}} \sum_{k\geq 0} \mathrm{e}(-m\,[\![\chi]\!]\cdot\infty)\,\mathrm{e}(-n\,[\![\chi^{-1}]\!]\cdot\infty)(4\pi^2\varrho\,[\![\chi]\!])^{k+s}m^{(k)}n^{(k+2s-1)} \end{aligned} \tag{3.4.26}$$

$$\mathrm{g}_{\Gamma,\mathsf{p}|\mathsf{q}}(-m,0,s) = \frac{1}{m^{1-s}}\frac{4\pi}{2s-1} \lim_{K\to\infty} \sum_{[\![\chi]\!] \in S_{\leq K}} \sum_{k\geq 0} \mathrm{e}(-m\,[\![\chi]\!]\cdot\infty)\varrho\,[\![\chi]\!]^s\,\frac{\pi^s}{\Gamma(s)} \tag{3.4.27}$$

$$\begin{aligned} &\mathrm{g}_{\Gamma,\mathsf{p}|\mathsf{q}}(-m,n,s) = \\ &\frac{1}{m^{1-s}n^s} \lim_{K\to\infty} \sum_{[\![\chi]\!] \in S_{\leq K}} \sum_{k\geq 0} \mathrm{e}(-m\,[\![\chi]\!]\cdot\infty)\,\mathrm{e}(n\,[\![\chi^{-1}]\!]\cdot\infty)(-1)^k(4\pi^2\varrho\,[\![\chi]\!])^{k+s}m^{(k)}n^{(k+2s-1)} \end{aligned} \tag{3.4.28}$$

Comparing with Lemma (3.2.6) we obtain a direct relationship between the coefficient functions $\mathrm{g}_{\Gamma,\mathsf{p}|\mathsf{q}}(m,n,s)$ arising from the Green's function, and the coefficient functions $\mathrm{c}^{\kappa}_{\Gamma,\mathsf{p}|\mathsf{q}}(m,n)$ of §3.2 which arise from the Rademacher sums.

Proposition 3.4.5. *Let Γ be a group commensurable with $G(\mathbb{Z})$, let $\mathsf{p},\mathsf{q} \in \mathsf{P}_\Gamma$, and let $\kappa > 0$*

and $m<0$. Then we have the following identities for $n>0$.

$$\mathrm{c}^{1-\kappa}_{\Gamma,\mathsf{p}|\mathsf{q}}(m,-n)=-m^{\kappa}n^{1-\kappa}\mathrm{g}_{\Gamma,\mathsf{p}|\mathsf{q}}(m,-n,\kappa) \tag{3.4.29}$$

$$\mathrm{c}^{1-\kappa}_{\Gamma,\mathsf{p}|\mathsf{q}}(m,0)=-m^{\kappa}\mathrm{g}_{\Gamma,\mathsf{p}|\mathsf{q}}(m,0,\kappa) \tag{3.4.30}$$

$$\mathrm{c}^{1-\kappa}_{\Gamma,\mathsf{p}|\mathsf{q}}(m,n)=-m^{\kappa}n^{1-\kappa}\mathrm{g}_{\Gamma,\mathsf{p}|\mathsf{q}}(m,n,\kappa) \tag{3.4.31}$$

Taking $\kappa=1$ in Proposition 3.4.5 we obtain the identity $\mathrm{c}_{\Gamma,\mathsf{p}|\mathsf{q}}(m,n)=-m\mathrm{g}_{\Gamma,\mathsf{p}|\mathsf{q}}(m,n,1)$ for $m<0$, which, by Theorem 3.3.5, the series expansion (3.4.25) and the identity $W(z,1)=\mathrm{e}(z)$, may be reformulated as

$$R^{(m)}_{\Gamma,\mathsf{p}|\mathsf{q}}(z)-\bar{R}^{(m)}_{\Gamma,\mathsf{p}|\mathsf{q}}(z)=-mG^{(m)}_{\Gamma,\mathsf{p}|\mathsf{q}}(z,1) \tag{3.4.32}$$

for $m<0$. The Γ^{q}-invariance of the difference $R^{(m)}_{\Gamma,\mathsf{p}|\mathsf{q}}(z)-\bar{R}^{(m)}_{\Gamma,\mathsf{p}|\mathsf{q}}(z)$ of classical and conjugate Rademacher sums of weight 0 now follows from the Γ^{q}-invariance of the Green's function coefficient $G^{(m)}_{\Gamma,\mathsf{p}|\mathsf{q}}(z,s)$ at $s=1$ (cf. [Nie73]). In particular, we have the following result.

Theorem 3.4.6. *Let Γ be a group commensurable with $G(\mathbb{Z})$ that has width one at infinity, let $\mathsf{p}\in\mathsf{P}_{\Gamma}$ be a cusp of Γ, and let $m<0$. Then the harmonic function $R^{(m)}_{\Gamma,\mathsf{p}}(z)-\bar{R}^{(m)}_{\Gamma,\mathsf{p}}(z)$ is Γ-invariant.*

4 Modified Rademacher sums

We have seen in §3.4 that the classical Rademacher sum $R^{\kappa(m)}_{\Gamma,\mathsf{p}}(z)$ of §3 is an automorphic integral for Γ only after the addition of a particular constant function, which does not generally vanish. In this section we introduce a modification of Rademacher's construction via which the correct constant term appears naturally. Our approach employs an analytic continuation of the component functions defining the classical Rademacher sums, and entails the assignment of a Dirichlet series to each triple $(\Gamma,\mathsf{p},\mathsf{q})$ where Γ is a group commensurable with the modular group and $\mathsf{p},\mathsf{q}\in\mathsf{P}_{\Gamma}$ are cusps for Γ.

4.1 Construction

For $\kappa,m\in\mathbb{Z}$ such that $\kappa\leq 0$ and $m<0$, and for $[\![\chi\in[\![G(\mathbb{Q})^{\times}$ we define the *continued Rademacher component function*, denoted $(z,s)\mapsto T^{\kappa(m)}_{[\![\chi}(z,s)$, by setting

$$T^{\kappa(m)}_{[\![\chi}(z,s)=\mathrm{e}(m\,[\![\chi\cdot z)\,\mathrm{t}^{\kappa}(m,[\![\chi\,,z,s)\,\mathrm{jac}([\![\chi\,,z)^{\kappa} \tag{4.1.1}$$

for $z \in \mathbb{H}$ and $s \in \mathbb{C}$ such that $\Re(s) > 1$ where $\mathrm{t}^\kappa(m, [\![\chi\,, z, s)$ is the *continued Rademacher regularization factor of weight* 2κ which is in turn given by

$$\mathrm{t}^\kappa(m, [\![\chi\,, z, s) = \Phi(s-2\kappa, 1+s-2\kappa, m\,[\![\chi\cdot\infty - m\,[\![\chi\cdot z)(2\pi\mathbf{i}(m\,[\![\chi\cdot z - m\,[\![\chi\cdot\infty))^{s-2\kappa} \tag{4.1.2}$$

in case $[\![\chi \in [\![G(\mathbb{Q})^\times$ and $\mathrm{t}^\kappa(m, [\![\chi\,, z, s) = 1$ otherwise. We also define a function $s \mapsto T^{\kappa(m)}_{0,[\![\chi}(s)$ by setting

$$\begin{aligned} T^{\kappa(m)}_{0,[\![\chi}(s) =& \,\mathrm{e}(m\,[\![\chi\cdot\infty)\,(2\pi\mathbf{i}(m\,[\![\chi\cdot 0 - m\,[\![\chi\cdot\infty))^{(s-2\kappa)}\,\mathrm{jac}([\![\chi\,, 0)^\kappa \\ &- \mathrm{e}(s/2)\,\mathrm{e}(m\,[\![\chi\cdot\infty)\,(2\pi\mathbf{i}(m\,[\![\chi\cdot\infty - m\,[\![\chi\cdot 0))^{(s-2\kappa)}\,\mathrm{jac}([\![\chi\,, 0)^\kappa \end{aligned} \tag{4.1.3}$$

in case $[\![\chi \in [\![G(\mathbb{Q})^{\times\times}$, and by setting $T^{\kappa(m)}_{0,[\![\chi}(s) = 0$ otherwise, and we define the *modified continued Rademacher component function*, denoted $(z,s) \mapsto Q^{\kappa(m)}_{[\![\chi}(z,s)$, by subtracting $T^{\kappa(m)}_{0,[\![\chi}(z,s)$ from $T^{\kappa(m)}_{[\![\chi}(z,s)$.

$$Q^{\kappa(m)}_{[\![\chi}(z,s) = T^{\kappa(m)}_{[\![\chi}(z,s) - T^{\kappa(m)}_{0,[\![\chi}(z,s) \tag{4.1.4}$$

Now for $U \subset [\![G(\mathbb{Q})$, and for $\kappa, m \in \mathbb{Z}$ such that $\kappa \leq 0$ and $m < 0$, we define the *continued Rademacher sum* $T^{\kappa(m)}_U(z,s)$, and the *modified continued Rademacher sum* $Q^{\kappa(m)}_U(z,s)$, by setting

$$T^{\kappa(m)}_U(z,s) = \sum_{[\![\chi\in U} T^{\kappa(m)}_{[\![\chi}(z,s), \tag{4.1.5}$$

$$Q^{\kappa(m)}_U(z,s) = \sum_{[\![\chi\in U} Q^{\kappa(m)}_{[\![\chi}(z,s), \tag{4.1.6}$$

and we define $T^{\kappa(m)}_U(z)$ and $Q^{\kappa(m)}_U(z)$ by taking the limit as s tends to 1 through the region $\Re(s) > 1$ in $T^{\kappa(m)}_U(z,s)$ and $Q^{\kappa(m)}_U(z,s)$, respectively, so long as these limits exist.

$$T^{\kappa(m)}_U(z) = \lim_{s\to 1^+} T^{\kappa(m)}_U(z,s) \tag{4.1.7}$$

$$Q^{\kappa(m)}_U(z) = \lim_{s\to 1^+} Q^{\kappa(m)}_U(z,s) \tag{4.1.8}$$

The function $Q^{\kappa(m)}_U(z)$ is the most important from the point of view of automorphy for subgroups of $G(\mathbb{Q})$, while the function $T^{\kappa(m)}_U(z)$ plays a special rôle in the critical case that $\kappa = 0$. We call $Q^{\kappa(m)}_U(z)$ the *modified Rademacher sum of weight* 2κ *and order* m *associated to* U, and we call $T^{\kappa(m)}_U(z)$ the *normalized Rademacher sum of weight* 2κ *and order* m *associated to* U.

Observe that we recover the Rademacher component function $R^{\kappa(m)}_{[\![\chi}(z)$ by taking $s = 1$

in either the continued Rademacher component function $T^{\kappa(m)}_{[\![\chi}(z,s)$ or the modified continued Rademacher component function $Q^{\kappa(m)}_{[\![\chi}(z,s)$.

Lemma 4.1.1. *Let $\kappa, m \in \mathbb{Z}$ such that $\kappa \leq 0$ and $m < 0$, and let $[\![\chi \in [\![G(\mathbb{Q})$. Then we have $T^{\kappa(m)}_{[\![\chi}(z,1) = R^{\kappa(m)}_{[\![\chi}(z)$ and $T^{\kappa(m)}_{0,[\![\chi}(1) = 0$, and hence $Q^{\kappa(m)}_{[\![\chi}(z,1) = R^{\kappa(m)}_{[\![\chi}(z)$.*

Despite the result of Lemma 4.1.1, it is generally not the case that the modified Rademacher sum $Q^{\kappa(m)}_U(z)$ and the classical Rademacher sum $R^{\kappa(m)}_U(z)$ coincide (cf. Proposition 4.4.10).

We typically take U to be of the form $U = [\![\Sigma_{\mathsf{p}}^{-1}\Sigma_{\mathsf{q}}$ where $\{\Sigma_{\mathsf{p}} \mid \mathsf{p} \in \mathsf{P}_\Gamma\}$ is a system of scaling cosets (cf. §2.6) for some group Γ commensurable with the modular group, and $\mathsf{p}, \mathsf{q} \in \mathsf{P}_\Gamma$ are cusps of Γ. In this case we write $Q^{\kappa(m)}_{\Gamma,\mathsf{p}|\mathsf{q}}(z)$ for $Q^{\kappa(m)}_U(z)$ and $T^{\kappa(m)}_{\Gamma,\mathsf{p}|\mathsf{q}}(z)$ for $T^{\kappa(m)}_U(z)$, suppressing the choice of scaling cosets from notation. A change in the choice of scaling cosets Σ_{p} and Σ_{q} replaces $Q^{\kappa(m)}_{\Gamma,\mathsf{p}|\mathsf{q}}(z)$ with a function of the form $\mathrm{e}(\alpha)Q^{\kappa(m)}_{\Gamma,\mathsf{p}|\mathsf{q}}(z+\beta)$ for some $\alpha, \beta \in \mathbb{Q}$, and similarly for $T^{\kappa(m)}_{\Gamma,\mathsf{p}|\mathsf{q}}(z)$.

In the case that p or q is the *infinite cusp* $\Gamma \cdot \infty$ we omit it from notation, writing $Q^{\kappa(m)}_{\Gamma,\mathsf{p}}(z)$ for $Q^{\kappa(m)}_{\Gamma,\mathsf{p}|\Gamma\cdot\infty}(z)$, and $Q^{\kappa(m)}_{\Gamma|\mathsf{q}}(z)$ for $Q^{\kappa(m)}_{\Gamma,\Gamma\cdot\infty|\mathsf{q}}(z)$, and similarly for the functions $T^{\kappa(m)}_{\Gamma,\mathsf{p}|\mathsf{q}}(z)$. The functions $Q^{\kappa(m)}_{\Gamma,\mathsf{p}}(z)$ and $T^{\kappa(m)}_{\Gamma,\mathsf{p}}(z)$ are the most important, for we shall see in §4.5 that $Q^{\kappa(m)}_{\Gamma,\mathsf{p}}(z)$ is an automorphic integral of weight 2κ for Γ with a single pole at the cusp p in case Γ has width one at infinity, and we will see in §4.4 that $Q^{\kappa(m)}_{\Gamma,\mathsf{p}}(z)$ and $T^{\kappa(m)}_{\Gamma,\mathsf{p}}(z)$ differ only by a constant function. We call $Q^{\kappa(m)}_{\Gamma,\mathsf{p}}(z)$ the *modified Rademacher sum of weight* 2κ *and order* m *associated to* Γ *at the cusp* p, and we call $T^{\kappa(m)}_{\Gamma,\mathsf{p}}(z)$ the *normalized Rademacher sum of weight* 2κ *and order* m *associated to* Γ *at the cusp* p. We shall see also in §4.5 that the Fourier expansion of the function $Q^{\kappa(m)}_{\Gamma,\mathsf{p}|\mathsf{q}}(z)$ is the expansion of $Q^{\kappa(m)}_{\Gamma,\mathsf{p}}(z)$ at the cusp q, in the sense of §2.7.

Just as for the classical Rademacher sums (cf. §3.1) we have the result that every modified Rademacher sum $Q^{\kappa(m)}_{\Gamma,\mathsf{p}|\mathsf{q}}(z)$ may be expressed in the form $Q^{\kappa(m)}_{\Gamma',\mathsf{p}'}(z)$ for some group Γ' with width one at infinity, and some cusp p' of Γ', and similarly for the normalized Rademacher sums $T^{\kappa(m)}_{\Gamma,\mathsf{p}|\mathsf{q}}(z)$. Precisely, we have

$$Q^{\kappa(m)}_{\Gamma,\mathsf{p}|\mathsf{q}}(z) = Q^{\kappa(m)}_{\Gamma^{\mathsf{q}},\mathsf{p}^{\mathsf{q}}}(z) \tag{4.1.9}$$

$$T^{\kappa(m)}_{\Gamma,\mathsf{p}|\mathsf{q}}(z) = T^{\kappa(m)}_{\Gamma^{\mathsf{q}},\mathsf{p}^{\mathsf{q}}}(z) \tag{4.1.10}$$

subject to the understanding that the data defining the right hand sides of (4.1.9) and (4.1.10) is related to the data defining the left hand sides of (4.1.9) and (4.1.10) by $\Gamma^{\mathsf{q}} = \Sigma_{\mathsf{q}}^{-1}\Sigma_{\mathsf{q}}$ and $\mathsf{p}^{\mathsf{q}} = \Sigma_{\mathsf{q}}^{-1} \cdot \mathsf{p}$ and $\Sigma_{\mathsf{p}^{\mathsf{q}}} = \Sigma_{\mathsf{q}}^{-1}\Sigma_{\mathsf{p}}$.

Our primary interest in this article is in the distinguished case that $\kappa = 0$. In order to simplify notation, and maintain consistency with the notation of §1, we write $Q^{(m)}_{\Gamma,\mathsf{p}|\mathsf{q}}(z)$ as a shorthand for $Q^{0(m)}_{\Gamma,\mathsf{p}|\mathsf{q}}(z)$, and we write $T^{(m)}_{\Gamma,\mathsf{p}|\mathsf{q}}(z)$ as a shorthand for $T^{0(m)}_{\Gamma,\mathsf{p}|\mathsf{q}}(z)$.

4.2 Coefficients

In order to recover explicit expressions of the Fourier coefficients of the modified Rademacher sums we employ the following generalizations of $\mathrm{Bl}^{\kappa}_{[\![\chi]\!]}(m,n)$ and $\mathrm{c}^{\kappa}_{S}(m,n)$ for $\kappa, m, n \in \mathbb{Z}$ such that $\kappa \leq 0$ and $m < 0 < n$. For $\kappa \leq 0$ and $m < 0 < n$ we define the entire function $s \mapsto \mathrm{Bl}^{\kappa}_{[\![\chi]\!]}(m,n,s)$ by setting

$$\mathrm{Bl}^{\kappa}_{[\![\chi]\!]}(m,n,s) = (-1)^{\kappa} \sum_{k \geq 0} (4\pi^2)^{k-\kappa+s} \varrho\,[\![\chi]\!]^{k-2\kappa+s}\, (-m)^{(k-2\kappa+s)} n^{(k+s-1)}, \tag{4.2.1}$$

so that $\mathrm{Bl}^{\kappa}_{[\![\chi]\!]}(m,n,1) = \mathrm{Bl}^{\kappa}_{[\![\chi]\!]}(m,n)$. We then define the *continued coefficient function* $s \mapsto \mathrm{c}^{\kappa}_{S}(m,n,s)$ for $\kappa \leq 0$ and $m < 0 < n$ by setting

$$\mathrm{c}^{\kappa}_{S}(m,n,s) = \sum_{[\![\chi]\!] \in S} \mathrm{Kl}^{\kappa}_{[\![\chi]\!]}(m,n)\, \mathrm{Bl}^{\kappa}_{[\![\chi]\!]}(m,n,s). \tag{4.2.2}$$

Lemma 4.2.1. *The functions $s \mapsto \mathrm{c}^{\kappa}_{S}(m,n,s)$ admit the following series representation.*

$$\begin{aligned} &\mathrm{c}^{\kappa}_{S}(m,n,s) \\ &= (-1)^{\kappa} \lim_{K \to \infty} \sum_{[\![\chi]\!] \in S_{\leq K}} \sum_{k \geq 0} \mathrm{e}(m\,[\![\chi]\!] \cdot \infty)\, \mathrm{e}(-n\,[\![\chi^{-1}]\!] \cdot \infty)(4\pi^2 \varrho\,[\![\chi]\!])^{k-\kappa+s}(-m)^{(k-2\kappa+s)} n^{(k+s-1)} \end{aligned} \tag{4.2.3}$$

We have $\mathrm{c}^{\kappa}_{S}(m,n,1) = \mathrm{c}^{\kappa}_{S}(m,n)$.

We usually take S to be of the form $S = [\![\Sigma_{\mathsf{p}}^{-1}\Sigma_{\mathsf{q}}]\!]^{\times}$ where Σ_{p} and Σ_{q} are scaling cosets (cf. §2.6) for a group Γ commensurable with $G(\mathbb{Z})$ at cusps $\mathsf{p}, \mathsf{q} \in \mathsf{P}_{\Gamma}$. We write $\mathrm{c}^{\kappa}_{\Gamma,\mathsf{p}|\mathsf{q}}(m,n,s)$ for $\mathrm{c}^{\kappa}_{S}(m,n,s)$ in the case that $S = [\![\Sigma_{\mathsf{p}}^{-1}\Sigma_{\mathsf{q}}]\!]^{\times}$.

Utilizing the continued coefficient functions $s \mapsto \mathrm{c}^{\kappa}_{S}(m,n,s)$ we generalize the function $F_S^{\kappa(m)}(z)_{\mathrm{van}}$ in case $\kappa \leq 0$ and $m < 0$ by setting

$$F_S^{\kappa(m)}(z,s)_{\mathrm{van}} = \sum_{n>0} \mathrm{c}^{\kappa}_{S}(m,n,s)\, \mathrm{e}(nz). \tag{4.2.4}$$

We write $F^{\kappa(m)}_{\Gamma,\mathsf{p}|\mathsf{q}}(z,s)_{\mathrm{van}}$ for $F_S^{\kappa(m)}(z,s)_{\mathrm{van}}$ in the case that $S = [\![\Sigma_{\mathsf{p}}^{-1}\Sigma_{\mathsf{q}}]\!]^{\times}$ for Γ a group commensurable with $G(\mathbb{Z})$ and for $\Sigma_{\mathsf{p}}, \Sigma_{\mathsf{q}} \in \Gamma\backslash G(\mathbb{Q})$ scaling cosets for Γ at cusps $\mathsf{p}, \mathsf{q} \in \mathsf{P}_{\Gamma}$. We write $F_S^{(m)}(z,s)_{\mathrm{van}}$ as a shorthand for $F_S^{0(m)}(z,s)_{\mathrm{van}}$, and interpret the notation $F^{(m)}_{\Gamma,\mathsf{p}|\mathsf{q}}(z,s)_{\mathrm{van}}$ similarly.

We have the following analogue of Proposition 3.2.10 for the functions $F_S^{\kappa(m)}(z,s)_{\mathrm{van}}$ in case S is of the form $[\![\Sigma_{\mathsf{p}}^{-1}\Sigma_{\mathsf{q}}]\!]^{\times}$.

Proposition 4.2.2. *Let Γ be a group commensurable with $G(\mathbb{Z})$, and let $\mathsf{p}, \mathsf{q} \in \mathsf{P}_{\Gamma}$ be cusps of*

Γ. *Let $\kappa, m \in \mathbb{Z}$ such that $\kappa \leq 0$ and $m < 0$. Then the series*

$$F^{\kappa(m)}_{\Gamma,\mathfrak{p}|\mathfrak{q}}(z,s)_{\mathrm{van}} = \sum_{n>0} \mathrm{c}^{\kappa}_{\Gamma,\mathfrak{p}|\mathfrak{q}}(m,n,s)\,\mathrm{e}(nz) \tag{4.2.5}$$

converges absolutely and locally uniformly in z and s for $z \in \mathbb{H}$ and $\Re(s) \geq 1$. In particular, the assignment $z \mapsto F^{\kappa(m)}_{\Gamma,\mathfrak{p}|\mathfrak{q}}(z,s)_{\mathrm{van}}$ is a holomorphic function on $\mathbb{H}$ whenever $\kappa \leq 0$ and $m < 0$ and $\Re(s) \geq 1$.

4.3 Dirichlet series

Given $[\![\chi]\!] \in [\![G(\mathbb{Q})]\!]^{\times}$, and $\kappa, m \in \mathbb{Z}$ such that $\kappa \leq 0$ and $m < 0$, we define a Dirichlet series $Z^{\kappa(m)}_{[\![\chi]\!]}(s)$ by setting

$$\begin{aligned} Z^{\kappa(m)}_{[\![\chi]\!]}(s) &= (-1)^{1-\kappa} \sum_{n\in\mathbb{Z}^{\times}} \mathrm{e}(m\,[\![\chi]\!]\cdot\infty)\,\mathrm{e}(-n\,[\![\chi^{-1}]\!]\cdot\infty)(4\pi^2\varrho\,[\![\chi]\!])^{1-s-\kappa}|m|^{(1-s-2\kappa)}n^{(-s)} \\ &= (-1)^{1-\kappa}\,\mathrm{e}(m\,[\![\chi]\!]\cdot\infty)(4\pi^2\varrho\,[\![\chi]\!])^{1-s-\kappa}|m|^{(1-s-2\kappa)} \sum_{n\in\mathbb{Z}^{\times}} \mathrm{e}(-n\,[\![\chi^{-1}]\!]\cdot\infty)n^{(-s)}. \end{aligned} \tag{4.3.1}$$

(Recall that $\mathbb{Z}^{\times}$ denotes the non-zero elements of $\mathbb{Z}$.) This series (4.3.1) converges absolutely and locally uniformly in the half plane $\Re(s) > 1$, and admits a meromorphic continuation to all of $\mathbb{C}$. Given $S \subset [\![G(\mathbb{Q})]\!]^{\times}$ we define the *zeta function of weight 2κ and order m associated to S*, to be denoted $Z^{\kappa(m)}_{S}(s)$, by setting

$$Z^{\kappa(m)}_{S}(s) = \sum_{[\![\chi]\!]\in S} Z^{\kappa(m)}_{[\![\chi]\!]}(s). \tag{4.3.2}$$

As it is defined here $Z^{\kappa(m)}_{S}$ is only a formal sum but it will be shown to converge, and admit an analytic continuation, for various choices of S in §4.4 (cf. Proposition 4.4.4, Theorem 4.4.5). In order to describe the analytic continuation of the zeta function $s \mapsto Z^{\kappa(m)}_{S}(s)$ explicitly, we formulate an identity which expresses it in terms of the Hurwitz zeta function. In preparation for this we define a function $s \mapsto \tilde{Z}^{\kappa(m)}_{[\![\chi]\!]}(s)$, for each $[\![\chi]\!] \in [\![G(\mathbb{Q})]\!]^{\times}$, by setting

$$\tilde{Z}^{\kappa(m)}_{[\![\chi]\!]}(s) = \mathrm{e}(m\,[\![\chi]\!]\cdot\infty)\varrho\,[\![\chi]\!]^{s-\kappa}\,(-2\pi\mathbf{i}m)^{(s-2\kappa)}(\mathrm{e}(s/2)-\mathrm{e}(-s/2))\zeta(1-\alpha_{[\![\chi]\!]},s) \tag{4.3.3}$$

where $\zeta(\alpha,s)$ is defined by $\zeta(\alpha,s) = \sum_{n\geq 0}(n+\alpha)^{-s}$ for $\Re(\alpha) > 0$ and $\Re(s) > 1$, and $\alpha_{[\![\chi]\!]} \in \mathbb{Q}$ is chosen so that

$$0 \leq \alpha_{[\![\chi]\!]} < 1, \quad \alpha_{[\![\chi]\!]} + \mathbb{Z} = -\chi^{-1}\cdot\infty + \mathbb{Z}. \tag{4.3.4}$$

We then define $\tilde{Z}_S^{\kappa(m)}(s)$ by setting $\tilde{Z}_S^{\kappa(m)}(s) = \sum_{[\![\chi]\!]\in S} \tilde{Z}_{[\![\chi]\!]}^{\kappa(m)}(s)$.

Lemma 4.3.1. *We have* $Z_S^{\kappa(m)}(1-s) = \tilde{Z}_S^{\kappa(m)}(s)$.

Proof. Let $[\![\chi]\!] \in [\![G(\mathbb{Q})]\!]^\times$ and choose $\alpha_{[\![\chi]\!]} \in \mathbb{Q}$ such that $0 \leq \alpha_{[\![\chi]\!]} < 1$ and $\alpha_{[\![\chi]\!]} + \mathbb{Z} = -\chi^{-1}\cdot\infty + \mathbb{Z}$ for any $\chi \in [\![\chi]\!]$. We will show that $Z_{[\![\chi]\!]}^{\kappa(m)}(s) = \tilde{Z}_{[\![\chi]\!]}^{\kappa(m)}(1-s)$. Observe that we have

$$\sum_{n\in\mathbb{Z}^\times} \mathrm{e}(-n\,[\![\chi^{-1}]\!]\cdot\infty)n^{(-s)} = \frac{1}{\Gamma(1-s)}\left(F\left(-[\![\chi^{-1}]\!]\cdot\infty, s\right) + \mathrm{e}(-s/2)F\left([\![\chi^{-1}]\!]\cdot\infty, s\right)\right), \tag{4.3.5}$$

where $F(\alpha, s)$ denotes the periodic zeta function, defined by $F(\alpha,s) = \sum_{n>0}\mathrm{e}(n\alpha)n^{-s}$ for $\Re(s) > 1$. The *Hurwitz relation*, which may be stated in the form

$$F(\alpha,s) + \mathrm{e}(-s/2)F(-\alpha,s) = \frac{(-2\pi\mathbf{i})^s}{\Gamma(s)}\zeta(1-\alpha, 1-s) \tag{4.3.6}$$

(cf. [KR01]) implies the identity

$$\sum_{n\in\mathbb{Z}^\times} \mathrm{e}(-n\,[\![\chi^{-1}]\!]\cdot\infty)n^{(-s)} = \frac{(-2\pi\mathbf{i})^s}{\Gamma(1-s)\Gamma(s)}\zeta(1-\alpha_{[\![\chi]\!]}, 1-s), \tag{4.3.7}$$

so the expression (4.3.1) may be reformulated as

$$Z_{[\![\chi]\!]}^{\kappa(m)}(s) = (-1)^{1-\kappa}\,\mathrm{e}(m\,[\![\chi]\!]\cdot\infty)(4\pi^2\varrho\,[\![\chi]\!])^{1-s-\kappa}|m|^{(1-s-2\kappa)}\frac{(-2\pi\mathbf{i})^s}{\Gamma(1-s)\Gamma(s)}\zeta(1-\alpha_{[\![\chi]\!]}, 1-s). \tag{4.3.8}$$

We obtain the coincidence of the right hand side of (4.3.8) with $\tilde{Z}_{[\![\chi]\!]}^{\kappa(m)}(1-s)$ by applying the functional equation for the Gamma function, which may be expressed as

$$(\mathrm{e}(s/2) - \mathrm{e}(-s/2))\Gamma(s)\Gamma(1-s) = 2\pi\mathbf{i} \tag{4.3.9}$$

This completes the proof of the claim. □

We write $Z_{\Gamma,\mathsf{p}|\mathsf{q}}^{\kappa(m)}(s)$ for $Z_S^{\kappa(m)}(s)$ in the case that $S = [\![\Sigma_\mathsf{p}^{-1}\Sigma_\mathsf{q}]\!]^\times$ for Γ a group commensurable with $G(\mathbb{Z})$ and for $\Sigma_\mathsf{p}, \Sigma_\mathsf{q} \in \Gamma\backslash G(\mathbb{Q})$ scaling cosets for Γ at cusps $\mathsf{p}, \mathsf{q} \in \mathsf{P}_\Gamma$.

4.4 Convergence

Our main objective in this section is to furnish explicit expressions for the Fourier expansions of the functions $z \mapsto T_{\Gamma,\mathsf{p}|\mathsf{q}}^{\kappa(m)}(z,s)$ and $z \mapsto Q_{\Gamma,\mathsf{p}|\mathsf{q}}^{\kappa(m)}(z,s)$, defined by the continued Rademacher sums

and modified continued Rademacher sums, respectively. This will yield the Fourier expansions of the modified Rademacher sums $Q^{\kappa(m)}_{\Gamma,\mathsf{p}|\mathsf{q}}(z)$ and the normalized Rademacher sums $T^{\kappa(m)}_{\Gamma,\mathsf{p}|\mathsf{q}}(z)$. In preparation for the derivation of these expansions we verify the absolute and locally uniform convergence of the series defining $T^{\kappa(m)}_{\Gamma,\mathsf{p}|\mathsf{q}}(z,s)$ and $Q^{\kappa(m)}_{\Gamma,\mathsf{p}|\mathsf{q}}(z,s)$ in case $\Re(s)>1$.

Proposition 4.4.1. *Let Γ be a group commensurable with $G(\mathbb{Z})$ and let $\mathsf{p},\mathsf{q}\in\mathsf{P}_\Gamma$ be cusps of Γ. Let $\kappa,m\in\mathbb{Z}$ such that $\kappa\leq 0$ and $m<0$. Then the continued Rademacher sum*

$$T^{\kappa(m)}_{\Gamma,\mathsf{p}|\mathsf{q}}(z,s)=\sum_{[\![\chi\in[\![\Sigma_\mathsf{p}^{-1}\Sigma_\mathsf{q}}T^{\kappa(m)}_{[\![\chi}(z,s) \tag{4.4.1}$$

is absolutely convergent, locally uniformly in z and s, for $z\in\mathbb{H}$ and $\Re(s)>1$.

Proof. We begin by observing that we have

$$T^{\kappa(m)}_{[\![\chi}(z,s)=\mathrm{e}(m\,[\![\chi\cdot\infty)\,\mathrm{e}(m\,[\![\chi\cdot z-m\,[\![\chi\cdot\infty,s-2\kappa)\,\mathrm{jac}([\![\chi\,,z)^\kappa \tag{4.4.2}$$

in case $[\![\chi\in[\![G(\mathbb{Q})^\times$, and $T^{\kappa(m)}_{[\![\chi}(z,s)=\mathrm{e}(m\,[\![\chi\cdot z)\,\mathrm{jac}([\![\chi\,,z)^\kappa$ otherwise, where $\mathrm{e}(z,s)$ denotes the generalized exponential function of (2.1.6). This identity (4.4.2) follows from an application of the *Kummer transformation* $\Phi(a,b,z)=\mathrm{e}(z)\Phi(b-a,b,-z)$ to the expression (4.1.1) defining $T^{\kappa(m)}_{[\![\chi}(z,s)$. The absolute and locally uniform convergence of (4.4.1) can be established by implementing methods present in the proofs of Proposition 3.2.8 and Theorem 3.3.2. Namely, the right hand side of (4.4.1) is very similar to the quantity R_+ defined in (3.3.15) and we can rewrite it so as to arrive at an analogue of (3.3.17). Then the coefficients of $\mathrm{e}(nz)$ in the resulting sum can be estimated just as the values $\mathrm{c}^\kappa_S(m,n)$ are in (3.2.27) for $\kappa>1$. □

Proposition 4.4.2. *Let Γ be a group commensurable with $G(\mathbb{Z})$ and let $\mathsf{p},\mathsf{q}\in\mathsf{P}_\Gamma$ be cusps of Γ. Let $\kappa,m\in\mathbb{Z}$ such that $\kappa\leq 0$ and $m<0$. Then the sum*

$$\sum_{[\![\chi\in[\![\Sigma_\mathsf{p}^{-1}\Sigma_\mathsf{q}}T^{\kappa(m)}_{0,[\![\chi}(s) \tag{4.4.3}$$

is absolutely convergent, locally uniformly in s, for $\Re(s)>1$.

Propositions 4.4.1 and 4.4.2 imply the absolute and locally uniform convergence of the expression defining the modified continued Rademacher sum $Q^{\kappa(m)}_{\Gamma,\mathsf{p}|\mathsf{q}}(z,s)$ for $z\in\mathbb{H}$ and $\Re(s)>1$.

Proposition 4.4.3. *Let Γ be a group commensurable with $G(\mathbb{Z})$, let $\mathsf{p},\mathsf{q}\in\mathsf{P}_\Gamma$ be cusps of Γ, and let Σ_p and Σ_q be scaling cosets for Γ at p and q, respectively. Let $\kappa\in\mathbb{Z}$ such that $\kappa\leq 0$,*

and let $m \in \mathbb{Z}^+$. Then the modified Rademacher sum

$$Q^{\kappa(m)}_{\Gamma,\mathsf{p}|\mathsf{q}}(z,s) = \sum_{[\![\chi \in [\![\Sigma_\mathsf{p}^{-1}\Sigma_\mathsf{q}} T^{\kappa(m)}_{[\![\chi}(z,s) - T^{\kappa(m)}_{0,[\![\chi}(s) \tag{4.4.4}$$

converges absolutely, and locally uniformly in z and s, for $z \in \mathbb{H}$ and $\Re(s) > 1$.

We next seek to relate the modified Rademacher sums $Q^{\kappa(m)}_{\Gamma,\mathsf{p}|\mathsf{q}}(z,s)$ to the functions $F^{\kappa(m)}_{\Gamma,\mathsf{p}|\mathsf{q}}(z,s)$ and $Z^{\kappa(m)}_{\Gamma,\mathsf{p}|\mathsf{q}}(s)$ of §4.2 and §4.3, respectively.

Proposition 4.4.4. *Let Γ be a group commensurable with $G(\mathbb{Z})$, let $\mathsf{p},\mathsf{q} \in \mathsf{P}_\Gamma$ be cusps of Γ, with scaling matrices Σ_p and Σ_q, respectively. Let $\kappa \in \mathbb{Z}$ such that $\kappa \leq 0$ and let $m \in \mathbb{Z}^+$. Then we have*

$$\sum_{[\![\chi \in [\![\Sigma_\mathsf{p}^{-1}\Sigma_\mathsf{q}} T^{\kappa(m)}_{0,[\![\chi}(s) = -Z^{\kappa(m)}_{\Gamma,\mathsf{p}|\mathsf{q}}(1-s) \tag{4.4.5}$$

for $\Re(s) > 1$ where the convergence is absolute and locally uniform.

Proof. Set $U = [\![\Sigma_\mathsf{p}^{-1}\Sigma_\mathsf{q}$ and $S = [\![\Sigma_\mathsf{p}^{-1}\Sigma_\mathsf{q}]\!]^\times$. Recall that $T^{\kappa(m)}_{0,[\![\chi}(s) = 0$ when $[\![\chi \notin [\![G(\mathbb{Q})^{\times\times}$, and otherwise

$$\begin{aligned} T^{\kappa(m)}_{0,[\![\chi}(s) = &\,\mathrm{e}(m\,[\![\chi\cdot\infty)\,(2\pi\mathbf{i}(m\,[\![\chi\cdot 0 - m\,[\![\chi\cdot\infty))^{(s-2\kappa)}\,\mathrm{jac}([\![\chi\,,0)^\kappa \\ &- \mathrm{e}(s/2)\,\mathrm{e}(m\,[\![\chi\cdot\infty)\,(2\pi\mathbf{i}(m\,[\![\chi\cdot\infty - m\,[\![\chi\cdot 0))^{(s-2\kappa)}\,\mathrm{jac}([\![\chi\,,0)^\kappa, \end{aligned} \tag{4.4.6}$$

by the definition (cf. §4.1) of $T^{\kappa(m)}_{0,[\![\chi}(s)$. We apply the identities $\chi\cdot\infty - \chi\cdot 0 = \varrho(\chi)/(-\chi^{-1}\cdot\infty)$ and $\mathrm{jac}([\![\chi\,,0) = \varrho(\chi)/(-\chi^{-1}\cdot\infty)^2$ (cf. §2.4) so as to write

$$T^{\kappa(m)}_{0,[\![\chi}(s) = \mathrm{e}(m\,[\![\chi]\!]\cdot\infty)\varrho\,[\![\chi]\!]^{s-\kappa}\,(-2\pi\mathbf{i}m)^{(s-2\kappa)}((-\,\chi^{-1}]\!]\cdot\infty)^{-s} - \mathrm{e}(s/2)(\chi^{-1}]\!]\cdot\infty)^{-s}). \tag{4.4.7}$$

Let $[\![\chi]\!] \in [\![G(\mathbb{Q})]\!]^\times$. Then we have

$$\sum_{[\![\chi \in [\![\chi]\!]} T^{\kappa(m)}_{0,[\![\chi}(s) = \mathrm{e}(m\,[\![\chi]\!]\cdot\infty)\varrho\,[\![\chi]\!]^{s-\kappa}\,(-2\pi\mathbf{i}m)^{(s-2\kappa)}\sum_{n\in\mathbb{Z}}{}^*((n+\alpha_{[\![\chi]\!]})^{-s} - \mathrm{e}(s/2)(n-\alpha_{[\![\chi]\!]})^{-s}) \tag{4.4.8}$$

where $\alpha_{[\![\chi]\!]} \in \mathbb{Q}$ is chosen so that $0 \leq \alpha_{[\![\chi]\!]} < 1$ and $\alpha_{[\![\chi]\!]} + \mathbb{Z} = -\chi^{-1}\cdot\infty + \mathbb{Z}$, and the superscript in the summation $\sum^*_{n\in\mathbb{Z}}$ indicates to omit the term corresponding to $n = 0$ in case $\alpha_{[\![\chi]\!]} = 0$.

We compute

$$\sum_{n\in\mathbb{Z}}^{*}((n+\alpha_{[\![\chi]\!]})^{-s}-\mathrm{e}(s/2)(n-\alpha_{[\![\chi]\!]})^{-s})=(\mathrm{e}(-s/2)-\mathrm{e}(s/2))\zeta(1-\alpha_{[\![\chi]\!]},s) \tag{4.4.9}$$

where $\zeta(\alpha,s)$ denotes the Hurwitz zeta function (cf. §4.3), so we have

$$\sum_{[\![\chi\in U^{\times\times}}T^{\kappa(m)}_{0,[\![\chi}(s)=\sum_{[\![\chi]\!]\in S^{\times}}\mathrm{e}(m\,[\![\chi]\!]\cdot\infty)\varrho\,[\![\chi]\!]^{s-\kappa}\,(-2\pi\mathbf{i}m)^{(s-2\kappa)}(\mathrm{e}(-s/2)-\mathrm{e}(s/2))\zeta(1-\alpha_{[\![\chi]\!]},s). \tag{4.4.10}$$

Comparing with the expression (4.3.3) defining $Z^{\kappa(m)}_{[\![\chi]\!]}(s)'$ we see that the right hand side of (4.4.10) coincides with $-Z^{\kappa(m)}_{S^{\times}}(s)'$. Now Lemma 4.3.1 yields the required identity (4.4.5). This completes the proof. □

Theorem 4.4.5. *Let Γ be a group commensurable with $G(\mathbb{Z})$ and let $\mathsf{p},\mathsf{q}\in\mathsf{P}_\Gamma$ be cusps of Γ. Let $\kappa,m\in\mathbb{Z}$ such that $\kappa\leq 0$ and $m<0$. Then we have*

$$Q^{\kappa(m)}_{\Gamma,\mathsf{p}|\mathsf{q}}(z,s)=\delta_{\Gamma,\mathsf{p}|\mathsf{q}}\,\mathrm{e}(-mz)+Z^{\kappa(m)}_{\Gamma,\mathsf{p}|\mathsf{q}}(1-s)+F^{\kappa(m)}_{\Gamma,\mathsf{p}|\mathsf{q}}(z,s)_{\mathrm{van}} \tag{4.4.11}$$

for $z\in\mathbb{H}$ and $\Re(s)>1$.

Proof. Let $m<0<\kappa$, let $s\in\mathbb{C}$ with $\Re(s)>1$, set $U=[\![\Sigma_{\mathsf{p}}^{-1}\Sigma_{\mathsf{q}}$ and set $S=[\![\Sigma_{\mathsf{p}}^{-1}\Sigma_{\mathsf{q}}]\!]^{\times}$. By the defining properties of scaling cosets (cf. §2.6) and by the definition (4.1.1) of the continued Rademacher component function $T^{1-\kappa(m)}_{[\![\chi}(z,s)$ we have

$$\sum_{[\![\chi\in U}T^{1-\kappa(m)}_{[\![\chi}(z,s)=\delta_{\Gamma,\mathsf{p}|\mathsf{q}}\,\mathrm{e}(mz)+\sum_{[\![\chi\in U^{\times}}T^{1-\kappa(m)}_{[\![\chi}(z,s) \tag{4.4.12}$$

so by Proposition 4.4.4, and the definition of the modified continued Rademacher component function $Q^{1-\kappa(m)}_{[\![\chi}(z,s)$, it suffices for us to show that

$$\sum_{[\![\chi\in U^{\times}}T^{1-\kappa(m)}_{[\![\chi}(z,s)=F^{1-\kappa(m)}_{S}(z,s)_{\mathrm{van}}. \tag{4.4.13}$$

We will verify the equality of (4.4.13) by using the Lipschitz summation formula (3.3.1) to transform the expression on the right hand side of (4.4.13) into that on the left. This is essentially the approach employed originally by Rademacher in [Rad39], and the reverse of that employed by Niebur in Lemma 4.2 of [Nie74], except that we can avoid the technical difficulties of [Rad39] and [Nie74], such as the need to employ the identity (3.3.10), by working only with the versions of the Lipschitz summation formula (viz., $\Re(s)>1$) in which both sides of the identity are

absolutely and locally uniformly convergent series.

We begin by inspecting Lemma 4.2.1 to find that

$$(-1)^{1-\kappa}\mathrm{c}_S^{1-\kappa}(m,n,s)$$
$$=\sum_{[\![\chi]\!]\in S}\sum_{k\geq 0}\mathrm{e}(m\,[\![\chi]\!]\cdot\infty)\,\mathrm{e}(-n\,[\![\chi^{-1}]\!]\cdot\infty)(4\pi^2\varrho\,[\![\chi]\!])^{k+\kappa+s-1}|m|^{(k+2\kappa+s-2)}n^{(k+s-1)}, \tag{4.4.14}$$

As a shorthand let us set $F_{\mathrm{van}}=F_S^{1-\kappa(m)}(z,s)_{\mathrm{van}}$. Then we have the following expression for F_{van} as a triple sum

$$(-1)^{1-\kappa}F_{\mathrm{van}}=$$
$$\sum_{n>0}\sum_{[\![\chi]\!]\in S}^{\lim}\sum_{k\geq 0}\mathrm{e}(m\,[\![\chi]\!]\cdot\infty)\,\mathrm{e}(-n\,[\![\chi^{-1}]\!]\cdot\infty)(4\pi^2\varrho\,[\![\chi]\!])^{k+\kappa+s-1}|m|^{(k+2\kappa+s-2)}n^{(k+s-1)}\,\mathrm{e}(nz) \tag{4.4.15}$$

where $\sum_{[\![\chi]\!]\in S}^{\lim}$ is a shorthand for $\lim_{K\to\infty}\sum_{[\![\chi]\!]\in S_{\leq K}}$. We now move the summation over n past the other two summations, simultaneously pulling the terms $\mathrm{e}(m\,[\![\chi]\!]\cdot\infty)$ outside the summation over k, and combining the exponents involving n. We thus obtain

$$(-1)^{1-\kappa}F_{\mathrm{van}}=$$
$$\sum_{[\![\chi]\!]\in S}^{\lim}\mathrm{e}(m\,[\![\chi]\!]\cdot\infty)\sum_{k\geq 0}(4\pi^2\varrho\,[\![\chi]\!])^{k+\kappa+s-1}|m|^{(k+2\kappa+s-2)}\sum_{n>0}n^{(k+s-1)}\,\mathrm{e}(n(z-[\![\chi^{-1}]\!]\cdot\infty)). \tag{4.4.16}$$

Applying the Lipschitz summation formula (3.3.1) to each summation over n in (4.4.16) we find that

$$(-1)^{1-\kappa}F_{\mathrm{van}}=$$
$$\sum_{[\![\chi]\!]\in S}^{\lim}\mathrm{e}(m\,[\![\chi]\!]\cdot\infty)\sum_{k\geq 0}(4\pi^2\varrho\,[\![\chi]\!])^{\kappa-1}|m|^{(k+2\kappa+s-2)}\sum_{n\in\mathbb{Z}}(2\pi\mathbf{i}\varrho\,[\![\chi]\!])^{k+s}(z+n-\chi^{-1}]\!]\cdot\infty)^{-k-s} \tag{4.4.17}$$

where $\chi^{-1}]\!]$ is any (left) coset of $B(\mathbb{Z})$ in the double coset $[\![\chi^{-1}]\!]$, for each $[\![\chi]\!]\in S$. Using the identity (2.4.5) we write F_{van} as $F_{\mathrm{van}}=\sum_{[\![\chi]\!]\in S}^{\lim}\mathrm{e}(m\,[\![\chi]\!]\cdot\infty)F_{\mathrm{van},[\![\chi]\!]}$ where

$$F_{\mathrm{van},[\![\chi]\!]}=\sum_{k\geq 0}\sum_{n\in\mathbb{Z}}(-2\pi\mathbf{i}\varrho\,[\![\chi]\!]\,m)^{k+2\kappa+s-2}(z+n-\chi^{-1}]\!]\cdot\infty)^{-k-2\kappa-s+2}\,\mathrm{jac}([\![\chi\,,z+n)^{1-\kappa}. \tag{4.4.18}$$

Recalling the definition of the generalized exponential function $\mathrm{e}(z,s)$ from (2.1.6) we see that

$$F_{\mathrm{van}} = \sum_{[\![\chi]\!]\in U^{\times}} \mathrm{e}(m\,[\![\chi]\!]\cdot\infty)\,\mathrm{e}\left(\frac{-\varrho\,[\![\chi]\!]\,m}{z-\chi^{-1}]\!]\cdot\infty}, s-2(1-\kappa)\right)\mathrm{jac}([\![\chi\ ,z)^{1-\kappa}. \tag{4.4.19}$$

According to (2.4.4) we have $\varrho\,[\![\chi]\!]\,/(z-\chi^{-1}\cdot\infty) = \chi\cdot\infty - \chi\cdot z$. Using this identity together with the Kummer transformation $\Phi(a,b,z) = \mathrm{e}(z)\Phi(b-a,b,-z)$ (cf. (2.1.3)) we readily find that

$$\mathrm{e}(m\,[\![\chi]\!]\cdot\infty)\,\mathrm{e}\left(\frac{-\varrho\,[\![\chi]\!]\,m}{z-\chi^{-1}]\!]\cdot\infty}, s-2(1-\kappa)\right) = \mathrm{e}(m\,[\![\chi]\!]\cdot z)\,\mathrm{t}^{1-\kappa}(m,[\![\chi\ ,z,s), \tag{4.4.20}$$

so that we have

$$F_{\mathrm{van}} = \sum_{[\![\chi]\!]\in U^{\times}} T_{[\![\chi]\!]}^{1-\kappa(m)}(z,s). \tag{4.4.21}$$

This establishes the required identity (4.4.13). □

In the course of proving Theorem 4.4.5 we have also established the Fourier expansion of the continued Rademacher sums $T_{\Gamma,\mathsf{p}|\mathsf{q}}^{\kappa(m)}(z,s)$. We record the result as follows.

Theorem 4.4.6. *Let Γ be a group commensurable with $G(\mathbb{Z})$ and let $\mathsf{p},\mathsf{q}\in\mathsf{P}_\Gamma$ be cusps of Γ. Let $\kappa,m\in\mathbb{Z}$ such that $\kappa\leq 0$ and $m<0$. Then we have*

$$T_{\Gamma,\mathsf{p}|\mathsf{q}}^{\kappa(m)}(z,s) = \delta_{\Gamma,\mathsf{p}|\mathsf{q}}\,\mathrm{e}(mz) + F_{\Gamma,\mathsf{p}|\mathsf{q}}^{\kappa(m)}(z,s)_{\mathrm{van}} \tag{4.4.22}$$

for $z\in\mathbb{H}$ and $\Re(s)>1$.

Theorem 4.4.5 will facilitate the identification of the Fourier expansion of the modified Rademacher sum $Q_{\Gamma,\mathsf{p}|\mathsf{q}}^{\kappa(m)}(z)$ obtained by sending s to 1 in $Q_{\Gamma,\mathsf{p}|\mathsf{q}}^{\kappa(m)}(z,s)$. Indeed, we will see presently (cf. Proposition 4.4.8) that the series $\tilde{F}_{\Gamma,\mathsf{p}|\mathsf{q}}^{\kappa(m)}(q)_{\mathrm{reg}}$ encodes the regular part of the fourier series expansion of $Q_{\Gamma,\mathsf{p}|\mathsf{q}}^{\kappa(m)}(z)$ precisely, in the case that $\kappa\leq 0$. In preparation for this we state the following result which identifies the value of the zeta function $Z_{\Gamma,\mathsf{p}|\mathsf{q}}^{\kappa(m)}(s)$ at $s=0$.

Proposition 4.4.7. *Let $S\subset[\![G(\mathbb{Q})]\!]^{\times}$, let $\kappa,m\in\mathbb{Z}$ such that $\kappa\leq 0$ and $m<0$. Then we have*

$$Z_S^{\kappa(m)}(0) = \mathrm{c}_S^{\kappa}(m,0). \tag{4.4.23}$$

Proof. According to the proof of Lemma 4.3.1 the function $\Gamma(s)Z_{[\![\chi]\!]}^{\kappa(m)}(s)$ coincides with

$$(-1)^{1-\kappa}\,\mathrm{e}(m\,[\![\chi]\!]\cdot\infty)(4\pi^2\varrho\,[\![\chi]\!])^{1-s-\kappa}|m|^{(1-s-2\kappa)}(F(\alpha_{[\![\chi]\!]},s)+\mathrm{e}(-s/2)F(-\alpha_{[\![\chi]\!]},s)) \tag{4.4.24}$$

where $0 \leq \alpha_{[\![\chi]\!]} < 1$ and $\alpha_{[\![\chi]\!]} + \mathbb{Z} = -\chi^{-1} \cdot \infty + \mathbb{Z}$. We have $F(\alpha, 0) + F(-\alpha, 0) = -1$ for $\alpha \in \mathbb{R}$ (cf. [KR01]), so taking $s = 0$ in (4.4.24) we obtain

$$Z^{\kappa(m)}_{[\![\chi]\!]}(0) = \mathrm{e}(m\,[\![\chi]\!] \cdot \infty)(-4\pi^2 \varrho\,[\![\chi]\!])^{1-\kappa} m^{(1-2\kappa)}. \tag{4.4.25}$$

Upon inspection of (3.2.21) we see that

$$\mathrm{c}^{\kappa}_{S}(m, 0) = \sum_{[\![\chi]\!] \in S} \mathrm{e}(m\,[\![\chi]\!] \cdot \infty)(-4\pi^2 \varrho\,[\![\chi]\!])^{1-\kappa} m^{(1-2\kappa)} \tag{4.4.26}$$

for $\kappa \leq 0$, and the claim now follows since $Z^{\kappa(m)}_{S}(s) = \sum_{[\![\chi]\!] \in S} Z^{\kappa(m)}_{[\![\chi]\!]}(s)$ by definition. □

Proposition 4.4.8. *Let Γ be a group commensurable with $G(\mathbb{Z})$, and let $\mathsf{p}, \mathsf{q} \in \mathsf{P}_\Gamma$ be cusps of Γ. Let $\kappa, m \in \mathbb{Z}$ such that $\kappa \leq 0$ and $m < 0$. Then we have*

$$Q^{\kappa(m)}_{\Gamma,\mathsf{p}|\mathsf{q}}(z) = \delta_{\Gamma,\mathsf{p}|\mathsf{q}}\,\mathrm{e}(mz) + F^{\kappa(m)}_{\Gamma,\mathsf{p}|\mathsf{q}}(z)_{\mathrm{reg}} \tag{4.4.27}$$

for $z \in \mathbb{H}$.

Proof. We take the limit as s tends to 1 in the right hand side of the identity (4.4.11). We have $F^{\kappa(m)}_{\Gamma,\mathsf{p}|\mathsf{q}}(z, 1)_{\mathrm{van}} = F^{\kappa(m)}_{\Gamma,\mathsf{p}|\mathsf{q}}(z)_{\mathrm{van}}$ by definition (cf. §4.2), and we have $Z^{\kappa(m)}_{\Gamma,\mathsf{p}|\mathsf{q}}(0) = \mathrm{c}^{\kappa}_{\Gamma,\mathsf{p}|\mathsf{q}}(m, 0)$ by Proposition 4.4.7. The required identity (4.4.27) now follows from the fact that $F^{\kappa(m)}_{\Gamma,\mathsf{p}|\mathsf{q}}(z)_{\mathrm{reg}} = \mathrm{c}^{\kappa}_{\Gamma,\mathsf{p}|\mathsf{q}}(m, 0) + F^{\kappa(m)}_{\Gamma,\mathsf{p}|\mathsf{q}}(z)_{\mathrm{van}}$, also by definition (cf. §3.2). □

The proof of Proposition 4.4.8 implies a result analogous to (4.4.27) for the normalized Rademacher sum $T^{\kappa(m)}_{\Gamma,\mathsf{p}|\mathsf{q}}(z)$.

Proposition 4.4.9. *Let Γ be a group commensurable with $G(\mathbb{Z})$ and let $\mathsf{p}, \mathsf{q} \in \mathsf{P}_\Gamma$ be cusps of Γ. Let $\kappa, m \in \mathbb{Z}$ such that $\kappa \leq 0$ and $m < 0$. Then we have*

$$T^{\kappa(m)}_{\Gamma,\mathsf{p}|\mathsf{q}}(z) = \delta_{\Gamma,\mathsf{p}|\mathsf{q}}\,\mathrm{e}(mz) + F^{\kappa(m)}_{\Gamma,\mathsf{p}|\mathsf{q}}(z)_{\mathrm{van}} \tag{4.4.28}$$

for $z \in \mathbb{H}$.

We may compare the functions $Q^{\kappa(m)}_{\Gamma,\mathsf{p}|\mathsf{q}}(z)$ and $R^{\kappa(m)}_{\Gamma,\mathsf{p}|\mathsf{q}}(z)$ defined by the modified and classical Rademacher sums, respectively. Inspecting Theorem 3.3.2 and Proposition 4.4.8 we obtain the precise relationship, which we record in the following proposition.

Proposition 4.4.10. *Let Γ be a group commensurable with $G(\mathbb{Z})$ and let $\mathsf{p}, \mathsf{q} \in \mathsf{P}_\Gamma$ be cusps of Γ. Let $\kappa, m \in \mathbb{Z}$ such that $\kappa \leq 0$ and $m < 0$. Then we have*

$$Q^{\kappa(m)}_{\Gamma,\mathsf{p}|\mathsf{q}}(z) = R^{\kappa(m)}_{\Gamma,\mathsf{p}|\mathsf{q}}(z) + \frac{1}{2}\mathrm{c}^{\kappa}_{\Gamma,\mathsf{p}|\mathsf{q}}(m, 0). \tag{4.4.29}$$

According to Propositions 4.4.8 and 4.4.9 the modified Rademacher sum $Q^{\kappa(m)}_{\Gamma,\mathsf{p}|\mathsf{q}}(z)$ and the normalized Rademacher sum $T^{\kappa(m)}_{\Gamma,\mathsf{p}|\mathsf{q}}(z)$ also differ only by a constant function.

Proposition 4.4.11. *Let Γ be a group commensurable with $G(\mathbb{Z})$ and let $\mathsf{p},\mathsf{q} \in \mathsf{P}_\Gamma$ be cusps of Γ. Let $\kappa, m \in \mathbb{Z}$ such that $\kappa \leq 0$ and $m < 0$. Then we have*

$$Q^{\kappa(m)}_{\Gamma,\mathsf{p}|\mathsf{q}}(z) = T^{\kappa(m)}_{\Gamma,\mathsf{p}|\mathsf{q}}(z) + \mathrm{c}^{\kappa}_{\Gamma,\mathsf{p}|\mathsf{q}}(m,0). \tag{4.4.30}$$

4.5 Variance

In this section we examine how the modified Rademacher sum $Q^{\kappa(m)}_{\Gamma,\mathsf{p}}(z)$ transforms under the weight 2κ action of Γ. Combining Theorem 3.4.4 and Proposition 4.4.10 we obtain the result that the modified Rademacher sum $Q^{\kappa(m)}_{\Gamma,\mathsf{p}}(z)$ is an automorphic integral of weight 2κ for Γ in case Γ has width one at infinity.

Theorem 4.5.1. *Let Γ be a group commensurable with $G(\mathbb{Z})$ that has width one at infinity and let $\mathsf{p} \in \mathsf{P}_\Gamma$ be a cusp of Γ. Let $\kappa, m \in \mathbb{Z}$ such that $\kappa \leq 0$ and $m < 0$. Then the modified Rademacher sum $Q^{\kappa(m)}_{\Gamma,\mathsf{p}}(z)$ is an automorphic integral of weight 2κ for Γ, and for $\mathsf{q} \in \mathsf{P}_\Gamma$ another cusp of Γ the Fourier expansion of the function $Q^{\kappa(m)}_{\Gamma,\mathsf{p}|\mathsf{q}}(z)$ is the expansion of $Q^{\kappa(m)}_{\Gamma,\mathsf{p}}(z)$ at the cusp q.*

The proof of Theorem 3.4.4 yields the following explicit description of the associated cusp form map $I_\kappa(\Gamma) \to S_{1-\kappa}(\Gamma)$ on the subspace of $I_\kappa(\Gamma)$ spanned by the modified Rademacher sums $Q^{\kappa(m)}_{\Gamma,\mathsf{p}}(z)$.

Proposition 4.5.2. *Let Γ be a group commensurable with $G(\mathbb{Z})$ that has width one at infinity and let $\mathsf{p} \in \mathsf{P}_\Gamma$ be a cusp of Γ. Let $\kappa, m \in \mathbb{Z}$ such that $\kappa \leq 0 < m$. Then*

$$m^\kappa Q^{\kappa(-m)}_{\Gamma,\mathsf{p}} \mapsto m^{1-\kappa} P^{1-\kappa(m)}_{\Gamma,\mathsf{p}} \mapsto 0, \tag{4.5.1}$$

under the maps $I_\kappa(\Gamma) \mapsto S_{1-\kappa}(\Gamma)$ of §2.7.

We next seek to establish the utility of Proposition 4.5.2 by identifying the subspace of $I_\kappa(\Gamma)$ that is spanned by the modified Rademacher sums $Q^{\kappa(m)}_{\Gamma,\mathsf{p}}(z)$.

Theorem 4.5.3. *Let Γ be a group commensurable with $G(\mathbb{Z})$ that has width one at infinity, and let $\kappa \in \mathbb{Z}$ with $\kappa \leq 0$. Then the set $\{Q^{\kappa(-m)}_{\Gamma,\mathsf{p}}(z) \mid \mathsf{p} \in \mathsf{P}_\Gamma, m \in \mathbb{Z}^+\}$ is a basis for the space of automorphic integrals of weight 2κ for Γ in case $\kappa < 0$. When $\kappa = 0$ the set $\{Q^{(-m)}_{\Gamma,\mathsf{p}}(z) \mid \mathsf{p} \in \mathsf{P}_\Gamma, m \in \mathbb{Z}^+\}$ spans a subspace of $I_0(\Gamma)$ of codimension 1.*

Proof. For $m > 0$ the function $Q^{\kappa(-m)}_{\Gamma,\mathsf{p}}(z)$ has principal part $\mathrm{e}(-mz)$ at the cusp p, and vanishes at all the other cusps, so the collection $\{Q^{\kappa(-m)}_{\Gamma,\mathsf{p}}(z) \mid \mathsf{p} \in \mathsf{P}_\Gamma, m \in \mathbb{Z}^+\}$ is linearly independent. To show that it furnishes a basis we follow Niebur's proof of Theorem 3.3 in [Nie74], employing

Petersson's generalized Riemann–Roch Theorem to show that the dimension of the subspace of $I_\kappa(\Gamma)$ containing automorphic integrals with poles of degree not more than m say at p is bounded above by m less the dimension of the space $S_{1-\kappa}(\Gamma)$, for m sufficiently large. □

We write $I_0'(\Gamma)$ for the subspace of $I_0(\Gamma)$ spanned by the modified Rademacher sums $Q_{\Gamma,\mathsf{p}}^{(-m)}(z)$ of weight 0 for varying $\mathsf{p} \in \mathsf{P}_\Gamma$ and $m \in \mathbb{Z}^+$.

Theorem 4.5.4. *Let Γ be a group commensurable with $G(\mathbb{Z})$ that has width one at infinity. Then for any $\kappa \in \mathbb{Z}$ the sequence*

$$0 \to M_\kappa(\Gamma) \to I_\kappa(\Gamma) \to S_{1-\kappa}(\Gamma) \to 0 \tag{4.5.2}$$

is exact.

Proof. The second map of (4.5.2) is an inclusion by definition, and we observed already in §2.7 that $M_\kappa(\Gamma)$ is the kernel of the associated cusp form map $I_\kappa(\Gamma) \to S_{1-\kappa}(\Gamma)$ (cf. Lemma 2.7.1), so we require to show only that the second to last map of (4.5.2) is surjective. For this we recall the fact (cf. [Iwa97]) that the holomorphic Poincaré series $P_{\Gamma,\mathsf{p}}^{1-\kappa(m)}(z)$, with varying $m \in \mathbb{Z}^+$, span the space $S_{1-\kappa}(\Gamma)$ of cusp forms of weight $2-2\kappa$ for Γ whenever $\kappa \leq 0$. Indeed, one need not vary the cusp p in order to obtain a spanning set. Given an arbitrary cusp form $g \in S_{1-\kappa}(\Gamma)$ we may then write

$$g = \sum_{n>0} a_n P_\Gamma^{1-\kappa(n)} \tag{4.5.3}$$

for some $a_n \in \mathbb{C}$, with only finitely many a_n non-zero. Employing Proposition 4.5.2 we see that the automorphic integral

$$f = \sum_{n>0} n^{2\kappa-1} a_n Q_\Gamma^{\kappa(-n)} \tag{4.5.4}$$

is mapped to g by the associated cusp form map. We conclude that the associated cusp form map is surjective. This completes the proof. □

Theorem 4.5.3 and Proposition 4.5.2 together describe the associated cusp form map $I_\kappa(\Gamma) \to S_{1-\kappa}(\Gamma)$ explicitly, and thus we obtain a powerful criterion for determining when an automorphic integral $f \in I_\kappa(\Gamma)$ lies in the subspace $M_\kappa(\Gamma)$ of modular forms. For example, we have the following corollary, indicating exactly when the modified Rademacher sum $Q_{\Gamma,\mathsf{p}}^{\kappa(m)}(z)$ is a modular form for Γ in the sense of §2.7.

Corollary 4.5.5. *Let Γ be a group commensurable with $G(\mathbb{Z})$ that has width one at infinity and let $\mathsf{p} \in \mathsf{P}_\Gamma$ be a cusp of Γ. Let $\kappa, m \in \mathbb{Z}$ such that $\kappa \leq 0$ and $m < 0$. Then the modified*

Rademacher sum $Q_{\Gamma,\mathfrak{p}}^{\kappa(m)}(z)$ *is a modular form of weight* 2κ *for* Γ *if and only if the cusp form* $P_{\Gamma,\mathfrak{p}}^{1-\kappa(-m)}(z)$ *vanishes identically.*

5 Structural applications

We have seen already in §4 various applications of the modified Rademacher sums, and the continuation procedure introduced in §4.1, such as the basis theorem (Theorem 4.5.3) for automorphic integrals, and the explicit description (cf. Proposition 4.5.2) of the associated cusp form map $I_\kappa(\Gamma) \to S_{1-\kappa}(\Gamma)$. In this section we explore further consequences of the modified Rademacher sum construction for the structure of the spaces of automorphic integrals associated to groups commensurable with the modular group. Applications of the modified Rademacher sums to monstrous moonshine and quantum gravity will be developed in §§6,7.

5.1 Constants

Perhaps the most striking application of the continuation procedure of §4.1 is the correction of the constant terms appearing in the classical Rademacher sums. For negative weights, the constant term in the Fourier expansion of an automorphic integral $f \in I_\kappa(\Gamma)$ is determined by the automorphy condition (2.7.13). At weight 0 constant functions are themselves automorphic integrals, so the classical Rademacher sums $R_{\Gamma,\mathfrak{p}}^{(m)}(z)$ of weight 0 do not fail to be automorphic, and it is natural then to wonder about the significance of the constant term $\mathrm{c}_{\Gamma,\mathfrak{p}}(m,0)$ appearing in the Fourier expansion of the modified Rademacher sum $Q_{\Gamma,\mathfrak{p}}^{(m)}(z)$ of weight 0 associated to a group Γ at a cusp $\mathfrak{p} \in \mathsf{P}_\Gamma$.

Let $\mathfrak{G}$ denote the set of subgroups of $G(\mathbb{Q})$ that are commensurable with $G(\mathbb{Z})$. Recall from §2.7 that for $\Gamma \in \mathfrak{G}$ we write $M_0(\Gamma)$ for the space of holomorphic functions on $\mathbb{H}$ that are invariant for the natural action of Γ and are meromorphic at the cusps of Γ. Let $\mathfrak{M}_0$ be the union of the spaces $M_0(\Gamma)$ for $\Gamma \in \mathfrak{G}$.

$$\mathfrak{M}_0 = \bigcup_{\Gamma \in \mathfrak{G}} M_0(\Gamma) \tag{5.1.1}$$

Since $\mathfrak{G}$ is closed under intersections, the set $\mathfrak{M}_0$ is in fact a subalgebra of the space $\mathcal{O}(\mathbb{H})$ of holomorphic functions on $\mathbb{H}$. Since $G(\mathbb{Q})$ is the commensurator of $G(\mathbb{Z})$, the algebra $\mathfrak{M}_0$ is $G(\mathbb{Q})$-invariant. The constant functions evidently constitute a $G(\mathbb{Q})$-invariant subspace $\mathbb{C}1 \subset \mathfrak{M}_0$. We may enquire as to the $G(\mathbb{Q})$-module structure of the quotient $\mathfrak{M}_0/\mathbb{C}1$. According to [Nor84] this quotient is irreducible, and the natural map $\mathfrak{M}_0 \to \mathfrak{M}_0/\mathbb{C}1$ admits a section.

Theorem 5.1.1 ([Nor84])**.** *There exists a unique* $G(\mathbb{Q})$*-submodule* $\mathfrak{M}_0' \subset \mathfrak{M}_0$ *with the property*

that the natural inclusions $\mathbb{C}1, \mathfrak{M}_0' \subset \mathfrak{M}_0$ *induce an isomorphism*

$$\mathfrak{M}_0 \cong \mathbb{C}1 \oplus \mathfrak{M}_0' \tag{5.1.2}$$

of $G(\mathbb{Q})$*-modules.*

According to Theorem 5.1.1, for each $f \in \mathfrak{M}_0$ there is a unique $c(f) \in \mathbb{C}$ with the property that $f + c(f) \in \mathfrak{M}_0'$. The value $c(f)$ is called the *Rademacher constant of* f. The assignment $f \mapsto c(f)$ defines a linear function on $\mathfrak{M}_0$ with kernel $\mathfrak{M}_0'$ which we call the *Rademacher constant function*. We suggest that the Rademacher constant function may be generalized to automorphic integrals of weight 0 in the following way. Let $\mathfrak{I}_0$ be the union of the spaces $I_0(\Gamma)$ for $\Gamma \in \mathfrak{G}$.

$$\mathfrak{I}_0 = \bigcup_{\Gamma \in \mathfrak{G}} I_0(\Gamma) \tag{5.1.3}$$

Since $\mathfrak{G}$ is closed under intersections, the set $\mathfrak{I}_0$ is also a subalgebra of $\mathcal{O}(\mathbb{H})$, and since $G(\mathbb{Q})$ is the commensurator of $G(\mathbb{Z})$, the algebra $\mathfrak{I}_0$ is also stable under the action of $G(\mathbb{Q})$. The constant functions again constitute a $G(\mathbb{Q})$-invariant subspace $\mathbb{C}1 \subset \mathfrak{I}_0$. We conjecture that the natural analogue of Theorem 5.1.1 holds for $\mathfrak{I}_0$. Precisely, we conjecture that there is a unique $G(\mathbb{Q})$-submodule $\mathfrak{I}_0' \subset \mathfrak{I}_0$ with the property that the natural inclusions induce an isomorphism

$$\mathfrak{I}_0 \cong \mathbb{C}1 \oplus \mathfrak{I}_0' \tag{5.1.4}$$

of $G(\mathbb{Q})$-modules. Further, we conjecture that $\mathfrak{I}_0'$ is just the space spanned by the modified Rademacher sums $Q_{\Gamma,\mathsf{p}}^{\kappa(-m)}(z)$ for $\Gamma \in \mathfrak{G}$ and $\mathsf{p} \in \mathsf{P}_\Gamma$ and $m \in \mathbb{Z}^+$.

5.2 Inner products

Let Γ be a group commensurable with $G(\mathbb{Z})$ and let $\kappa \in \mathbb{Z}^+$. Then the space $S_\kappa(\Gamma)$ of cusp forms of weight 2κ for Γ becomes a Hilbert space when equipped with the *Petersson inner product*, defined by setting

$$\langle f, g \rangle = \int_{\mathfrak{F}_\Gamma} f(z)\overline{g(z)}\Im(z)^{2\kappa} \mathrm{d}\mu(z) \tag{5.2.1}$$

for $f, g \in S_\kappa(\Gamma)$, where $\mathfrak{F}_\Gamma$ is a fundamental domain for Γ. Let us define a *normalized Petersson inner product of weight* 2κ *for* Γ, to be denoted $\langle \cdot, \cdot \rangle_\Gamma^\kappa$, by setting

$$\langle f, g \rangle_\Gamma^\kappa = (4\pi)^{(2\kappa-2)} \int_{\mathfrak{F}_\Gamma} f(z)\overline{g(z)}\Im(z)^{2\kappa} \mathrm{d}\mu(z) \tag{5.2.2}$$

for $f, g \in S_\kappa(\Gamma)$. It is well known (and follows from Theorem 3.4.3), that for $\kappa, m \in \mathbb{Z}^+$ and $\mathfrak{p}$ a cusp for Γ, the holomorphic Poincaré series $P_{\Gamma,\mathfrak{p}}^{\kappa(m)}(z)$ is a cusp form of weight 2κ for Γ in case Γ has width one at infinity. A standard calculation (cf. [Iwa97]) shows that we have

$$\left\langle P_{\Gamma,\mathfrak{p}}^{\kappa(m)}, P_{\Gamma,\mathfrak{q}}^{\kappa(n)} \right\rangle_\Gamma^\kappa = n^{2-2\kappa}(\delta_{\Gamma,\mathfrak{p}|\mathfrak{q}}\delta_{m,n} + \mathrm{c}_{\Gamma,\mathfrak{p}|\mathfrak{q}}^\kappa(m,n)) \tag{5.2.3}$$

for $\mathfrak{p}, \mathfrak{q} \in \mathsf{P}_\Gamma$ and $m, n \in \mathbb{Z}^+$. According to Theorem 3.4.3 the right hand side of (5.2.3) is, up to the scalar factor $n^{2-2\kappa}$, the coefficient of $\mathrm{e}(nz)$ in the expansion $P_{\Gamma,\mathfrak{p}|\mathfrak{q}}^{\kappa(m)}(z)$ of the Poincaré series $P_{\Gamma,\mathfrak{p}}^{\kappa(m)}(z)$ at $\mathfrak{q}$. Now the modified Rademacher sums $Q_{\Gamma,\mathfrak{p}}^{\kappa(m)}(z)$, for $-\kappa, -m \in \mathbb{Z}^+$, span the space $I_\kappa(\Gamma)$ of automorphic integrals for weight 2κ for Γ, according to Theorem 4.5.3, and thus serve an analogous rôle for the spaces $I_\kappa(\Gamma)$ as the Poincaré series $P_{\Gamma,\mathfrak{p}}^{\kappa(m)}(z)$ do for the spaces $S_\kappa(\Gamma)$. It is natural then to use the right hand side of (5.2.3) to extend the normalized Petersson inner product to negative weights. In light of the fact, established in Proposition 4.4.8, that the coefficient of $\mathrm{e}(-nz)$ in the expansion $Q_{\Gamma,\mathfrak{p}|\mathfrak{q}}^{\kappa(-m)}(z)$ of the modified Rademacher sum $Q_{\Gamma,\mathfrak{p}}^{\kappa(-m)}(z)$ at $\mathfrak{q}$ is $\delta_{\Gamma,\mathfrak{p}|\mathfrak{q}}\delta_{m,n}$ for $m, n \in \mathbb{Z}^+$, we define an inner product $\langle \cdot, \cdot \rangle_\Gamma^\kappa$ on $I_\kappa(\Gamma)$ by setting

$$\left\langle Q_{\Gamma,\mathfrak{p}}^{\kappa(-m)}, Q_{\Gamma,\mathfrak{q}}^{\kappa(-n)} \right\rangle_\Gamma^\kappa = n^{2-2\kappa}\delta_{\Gamma,\mathfrak{p}|\mathfrak{q}}\delta_{m,n} \tag{5.2.4}$$

for $\mathfrak{p}, \mathfrak{q} \in \mathsf{P}_\Gamma$ and $m, n \in \mathbb{Z}^+$, when $\kappa < 0$. In case $\kappa = 0$ the modified Rademacher sums $Q_{\Gamma,\mathfrak{p}}^{\kappa(-m)}(z)$ span a subspace $I_0'(\Gamma)$ of $I_0(\Gamma)$ of codimension 1 (cf. Theorem 4.5.3 and §5.1) and we may use the formula (5.2.4) to define an inner product on this subspace $I_0'(\Gamma)$.

$$\left\langle Q_{\Gamma,\mathfrak{p}}^{(-m)}, Q_{\Gamma,\mathfrak{q}}^{(-n)} \right\rangle_\Gamma^0 = n^2\delta_{\Gamma,\mathfrak{p}|\mathfrak{q}}\delta_{m,n} \tag{5.2.5}$$

A complement to $I_0'(\Gamma)$ in $I_0(\Gamma)$ is spanned by the constant functions. We extend the inner product $\langle \cdot, \cdot \rangle_\Gamma^0$ to all of $I_0(\Gamma)$ by adopting the convention that

$$\left\langle Q_{\Gamma,\mathfrak{p}}^{(-m)}, f \right\rangle_\Gamma^0 = 0 \tag{5.2.6}$$

for all $\mathfrak{p} \in \mathsf{P}_\Gamma$ and $m \in \mathbb{Z}^+$ whenever f is identically constant.

5.3 Branching

Let Δ and Γ be groups commensurable with $G(\mathbb{Z})$ and suppose that Δ is a subgroup of Γ. Then an automorphic integral for Γ is also an automorphic integral for Δ. By Theorem 4.5.3 the modified Rademacher sums span the spaces of automorphic integrals, so it is natural to consider the problem of expressing the modified Rademacher sums $Q_{\Gamma,\mathfrak{p}}^{\kappa(m)}$ associated to Γ in terms of the modified Rademacher sums $Q_{\Delta,\mathfrak{r}}^{\kappa(n)}$ associated to Δ.

Suppose then that Δ and Γ are groups commensurable with $G(\mathbb{Z})$ and Δ is a subgroup of

Γ. Let p be a cusp for Γ, let Σ_{p} be a scaling coset for Γ at p, and choose a system $\{\Sigma_{\mathsf{r}} \mid \mathsf{r} \in \mathsf{P}_\Delta\}$ of scaling cosets for Δ. Suppose for now that κ is a negative integer. Then by Theorem 4.5.3 we have

$$Q_{\Gamma,\mathsf{p}}^{\kappa(m)}(z) = \sum_{\mathsf{r}\in\mathsf{P}_\Delta} b_{\mathsf{r}}^n Q_{\Delta,\mathsf{r}}^{\kappa(n)}(z) \tag{5.3.1}$$

for some $b_{\mathsf{r}}^n \in \mathbb{C}$, and these b_{r}^n are non-zero for only finitely many $n < 0$. Furthermore, since the modified Rademacher sum $Q_{\Gamma,\mathsf{p}}^{\kappa(m)}(z)$ has no poles away from p, the coefficient b_{r}^n can be non-zero only when r lies in the preimage of p under the natural map $\mathsf{X}_\Delta \to \mathsf{X}_\Gamma$, which is to say, b_{r}^n is non-zero only when $\mathsf{r} \in \Delta\backslash\mathsf{p}$. It is natural then to consider the map

$$\begin{aligned} \Delta\backslash\Sigma_{\mathsf{p}} &\to \Delta\backslash\mathsf{p} \\ \Delta\sigma &\mapsto \Delta\sigma\cdot\infty \end{aligned} \tag{5.3.2}$$

sending right cosets of Δ in Σ_{p} to cusps of Δ contained in $\mathsf{p} \subset \hat{\mathbb{Q}}$. The map (5.3.2) is always surjective. We may ask under what circumstances it is also injective.

Lemma 5.3.1. *Let Δ and Γ be groups commensurable with $G(\mathbb{Z})$ and suppose that Δ is a subgroup of Γ. Let p be a cusp for Γ and let Σ_{p} be a scaling coset for Γ at p. Then the map $\Delta\backslash\Sigma_{\mathsf{p}} \to \Delta\backslash\mathsf{p}$ of (5.3.2) is injective if and only if $\Delta\sigma$ is a scaling coset for Δ at $\Delta\sigma\cdot\infty$ for every $\Delta\sigma \in \Delta\backslash\Sigma_{\mathsf{p}}$.*

Proof. Suppose the map $\Delta\backslash\Sigma_{\mathsf{p}} \to \Delta\backslash\mathsf{p}$ of (5.3.2) is injective, and let $\sigma \in \Sigma_{\mathsf{p}}$. The forward implication of the lemma follows if we can show that $\Delta\sigma$ is a scaling coset for Δ at $\Delta\sigma\cdot\infty$. For this it suffices to show that $(\sigma^{-1}\Delta\sigma)_\infty = B(\mathbb{Z})$. Since $\sigma^{-1}\Delta\sigma$ is a subgroup of $\sigma^{-1}\Gamma\sigma$, and $(\sigma^{-1}\Gamma\sigma)_\infty = (\Sigma_{\mathsf{p}}^{-1}\Sigma_{\mathsf{p}})_\infty = B(\mathbb{Z})$ by the defining properties of Σ_{p}, we have the inclusion $(\sigma^{-1}\Delta\sigma)_\infty \subset B(\mathbb{Z})$, so it suffices to show that the translation T belongs to $\sigma^{-1}\Delta\sigma$. Now $\Delta\sigma T$ is also a coset of Δ in Σ_{p}, since $\Gamma\sigma = \Sigma_{\mathsf{p}}$ is a union of left cosets of $B(\mathbb{Z})$. Since T fixes ∞ we have $\Delta\sigma T\cdot\infty = \Delta\sigma\cdot\infty$, and thus the cosets $\Delta\sigma T$ and $\Delta\sigma$ coincide by the assumed injectivity of the map (5.3.2). It follows that $T \in \sigma^{-1}\Delta\sigma$, so that $\Delta\sigma$ is indeed a scaling coset for Δ.

For the reverse implication suppose that $\Delta\sigma$ is a scaling coset for Δ for every $\sigma \in \Sigma_{\mathsf{p}}$. Let $\sigma, \sigma' \in \Sigma_{\mathsf{p}}$ and suppose that $\Delta\sigma\cdot\infty = \Delta\sigma'\cdot\infty$. Then $\sigma^{-1}\delta\sigma' \in (\Sigma_{\mathsf{p}}^{-1}\Sigma_{\mathsf{p}})_\infty = B(\mathbb{Z})$ for some $\delta \in \Delta$, so $\sigma^{-1}\delta\sigma' = T^n$ for some $n \in \mathbb{Z}$, and this implies $\Delta\sigma' = \Delta\sigma T^n$. Now the assumption that $\Delta\sigma$ is a scaling coset for Δ implies that $\Delta\sigma$ is a union of left cosets of $B(\mathbb{Z})$, so $\Delta\sigma T^n = \Delta\sigma$. We conclude that $\Delta\sigma' = \Delta\sigma$, so that the map $\Delta\backslash\Sigma_{\mathsf{p}} \to \Delta\backslash\mathsf{p}$ is indeed injective. This completes the proof of the lemma. □

According to the discussion of §2.6 a coset $\Delta\sigma \in \Delta\backslash\Sigma_{\mathsf{p}}$ becomes a scaling coset for Δ at $\Delta\sigma\cdot\infty$ only once we multiply it on the right by $[\beta]$ (cf. (2.3.3)) for some $\beta \in \mathbb{Q}^+$. In fact,

this β is a positive integer, for we have $(\sigma^{-1}\Delta\sigma)_\infty \subset (\Sigma_\mathsf{p}^{-1}\Sigma_\mathsf{p})_\infty = B(\mathbb{Z})$, so that $(\sigma^{-1}\Delta\sigma)_\infty$ is generated by T^n for some $n \in \mathbb{Z}^+$. In order that $\Delta\sigma[\beta]$ be a scaling coset for Δ at $\Delta\sigma\cdot\infty$ we should have $([1/\beta]\sigma^{-1}\Delta\sigma[\beta])_\infty = B(\mathbb{Z})$, so the computation $[1/\beta]T^n[\beta] = T^{n/\beta}$ shows that $\beta = n$. Evidently we may define a function $\Delta\backslash\Sigma_\mathsf{p} \to \mathbb{Z}^+$ by mapping the coset $\Delta\sigma$ to the positive integer n such that T^n generates $(\sigma^{-1}\Delta\sigma)_\infty$. We next show that this function $\Delta\backslash\Sigma_\mathsf{p} \to \mathbb{Z}^+$ factors through the map $\Delta\backslash\Sigma_\mathsf{p} \to \Delta\backslash\mathsf{p}$ of (5.3.2).

Lemma 5.3.2. *Let Δ and Γ be groups commensurable with $G(\mathbb{Z})$, and suppose that Δ is a subgroup of Γ. Let $\mathsf{p} \in \mathsf{P}_\Gamma$ be a cusp of Γ and let Σ_p be a scaling coset for Γ at p. Then for any $\sigma_1, \sigma_2 \in \Sigma_\mathsf{p}$ the identity $\Delta\sigma_1\cdot\infty = \Delta\sigma_2\cdot\infty$ implies $(\sigma_1^{-1}\Delta\sigma_1)_\infty = (\sigma_2^{-1}\Delta\sigma_2)_\infty$.*

Proof. Set $\mathsf{r} = \Delta\sigma_1\cdot\infty = \Delta\sigma_2\cdot\infty$. Since $(\sigma^{-1}\Gamma\sigma)_\infty = (\Sigma_\mathsf{p}^{-1}\Sigma_\mathsf{p})_\infty = B(\mathbb{Z})$ for any $\sigma \in \Sigma_\mathsf{p}$ the groups $(\sigma_1^{-1}\Delta\sigma_1)_\infty$ and $(\sigma_2^{-1}\Delta\sigma_2)_\infty$ are both contained in $B(\mathbb{Z})$. Let n_1 and n_2 be the positive integers such that $(\sigma_1^{-1}\Delta\sigma_1)_\infty$ is generated by T^{n_1} and $(\sigma_2^{-1}\Delta\sigma_2)_\infty$ is generated by T^{n_2}. Then $\Delta\sigma_1[n_1]$ and $\Delta\sigma_2[n_2]$ are both scaling cosets for Δ at r, so $\Delta\sigma_1[n_1] = \Delta\sigma_2[n_2]T^\alpha$ for some $\alpha \in \mathbb{Q}$. This implies that $[n_2]T^\alpha[1/n_1]$ is an element of $(\Sigma_\mathsf{p}^{-1}\Sigma_\mathsf{p})_\infty = B(\mathbb{Z})$, so that $[n_2]T^\alpha[1/n_1] = T^k$ for some $k \in \mathbb{Z}$. Now we compute

$$[n_2]T^\alpha[1/n_1] = \begin{bmatrix} n_2 & 0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} 1 & \alpha \\ 0 & 1 \end{bmatrix}\begin{bmatrix} 1 & 0 \\ 0 & n_1 \end{bmatrix} = \begin{bmatrix} n_2 & n_2\alpha n_1 \\ 0 & n_1 \end{bmatrix} = \begin{bmatrix} 1 & k \\ 0 & 1 \end{bmatrix} \tag{5.3.3}$$

and conclude that $n_1 = n_2$, as we required to show. □

On the basis of Lemma 5.3.2 we may define a function $\Delta\backslash\mathsf{p} \to \mathbb{Z}^+$ for each cusp p of Γ by first choosing a scaling coset Σ_p for Γ at p, and then sending the cusp $\mathsf{r} \in \Delta\backslash\mathsf{p}$ of Δ to the unique positive integer n satisfying $(\sigma^{-1}\Delta\sigma)_\infty = \langle T^n\rangle$ when $\mathsf{r} = \Delta\sigma\cdot\infty$. Observe now that this map does not depend upon the choice of scaling coset Σ_p, for if we replace $\sigma \in \Sigma_\mathsf{p}$ with $\sigma' = \sigma T^\alpha \in \Sigma_\mathsf{p}T^\alpha$ for some $\alpha \in \mathbb{Q}$, then we have

$$((\sigma')^{-1}\Delta\sigma')_\infty = (T^{-\alpha}\sigma^{-1}\Delta\sigma T^\alpha)_\infty = T^{-\alpha}(\sigma^{-1}\Delta\sigma)_\infty T^\alpha = (\sigma^{-1}\Delta\sigma)_\infty. \tag{5.3.4}$$

In light of this we may define a map $w_\Gamma : \mathsf{P}_\Delta \to \mathbb{Z}^+$, which we call the *width function associated to* Γ, by setting $w_\Gamma(\mathsf{r}) = n$ in case $\mathsf{r} = \Delta\sigma\cdot\infty$ and T^n generates $(\sigma^{-1}\Delta\sigma)_\infty$ and $\Gamma\sigma$ is a scaling coset for Γ at the unique cusp of Γ containing r.

$$\begin{gathered} w_\Gamma : \mathsf{P}_\Delta \to \mathbb{Z}^+ \\ \mathsf{r} = \Delta\sigma\cdot\infty,\ \Gamma\sigma \in \mathfrak{S}_{\Gamma\mathsf{r}} \Longrightarrow (\sigma^{-1}\Delta\sigma)_\infty = \langle T^{w_\Gamma(\mathsf{r})}\rangle \end{gathered} \tag{5.3.5}$$

We call $w_\Gamma(\mathsf{r})$ the *width of* r *with respect to* Γ. The next result gives upper bounds for the width functions.

Lemma 5.3.3. *Let Δ and Γ be groups commensurable with $G(\mathbb{Z})$, and suppose that Δ is a subgroup of Γ. Then we have*

$$\sum_{\mathsf{r}\in\Delta\backslash\mathsf{p}} w_\Gamma(\mathsf{r}) = \#\Delta\backslash\Gamma \tag{5.3.6}$$

for every cusp p of Γ.

Proof. Let p be a cusp of Γ and choose a scaling coset Σ_p for Γ at p, and let us temporarily write c_p for the map $\Delta\sigma \mapsto \Delta\sigma\cdot\infty$ of (5.3.2). We claim that $w_\Gamma(\mathsf{r})$ is the cardinality of the preimage of r under c_p, for each $\mathsf{r}\in\Delta\backslash\mathsf{p}$. The identity (5.3.6) follows from this because the cardinality $\#\Delta\backslash\Sigma_\mathsf{p}$ of the source of c_p is just the number of cosets of Δ in Γ. To compute $\#c_\mathsf{p}^{-1}(\mathsf{r})$ suppose $\Delta\sigma\cdot\infty = \Delta\sigma'\cdot\infty$ for some $\sigma,\sigma'\in\Sigma_\mathsf{p}$. Then $\sigma^{-1}\delta\sigma'\in(\Sigma_\mathsf{p}^{-1}\Sigma_\mathsf{p})_\infty = B(\mathbb{Z})$ for some $\delta\in\Delta$, so that $\Delta\sigma' = \Delta\sigma T^k$ for some $k\in\mathbb{Z}$. Now $\Delta\sigma T$ is an element of $\Delta\backslash\Sigma_\mathsf{p}$ whenever $\Delta\sigma$ is, for Σ_p is a union of left cosets of $B(\mathbb{Z})$ by construction. So we have $c_\mathsf{p}^{-1}(\mathsf{r}) = \{\Delta\sigma T^k \mid k\in\mathbb{Z}\}$ for any $\sigma\in\Sigma_\mathsf{p}$ satisfying $\mathsf{r} = \Delta\sigma\cdot\infty$. If $w_\Gamma(\mathsf{r}) = n$, so that $(\sigma^{-1}\Delta\sigma)_\infty = \langle T^n\rangle$ then $\Delta\sigma T^k = \Delta\sigma T^l$ if and only if $n|(k-l)$. This establishes the claim that $\#c_\mathsf{p}^{-1}(\mathsf{r}) = w_\Gamma(\mathsf{r})$, and completes the proof. □

We can now compute the coefficients b_r^n in (5.3.1) explicitly, and thus obtain a branching theorem for the modified Rademacher sums. Even though we formulated the expression (5.3.1) under the assumption that κ be negative, our methods will apply for all $\kappa\in\mathbb{Z}$ such that $\kappa\leq 0$.

Theorem 5.3.4. *Let Δ and Γ be groups commensurable with $G(\mathbb{Z})$. Suppose that Δ is a subgroup of Γ, and suppose that Δ and Γ both have width one at infinity. Let p be a cusp of Γ and let Σ_p be a scaling coset for Γ at p. Let $\kappa, m\in\mathbb{Z}$ such that $\kappa\leq 0$ and $m<0$. Then there exists a system of scaling cosets $\{\Sigma_\mathsf{r} \mid \mathsf{r}\in\mathsf{P}_\Delta\}$ for Δ for which we have*

$$Q_{\Gamma,\mathsf{p}}^{\kappa(m)}(z) = \sum_{\mathsf{r}\in\Delta\backslash\mathsf{p}} w_\Gamma(\mathsf{r}) Q_{\Delta,\mathsf{r}}^{\kappa(mw_\Gamma(\mathsf{r}))}(z). \tag{5.3.7}$$

Proof. Let Δ and Γ be as in the statement of the lemma. Under the assumption that Γ has width one at infinity, and the convention that $\Sigma_{\Gamma\cdot\infty} = \Gamma$ in this case (cf. §2.6), we have

$$Q_{\Gamma,\mathsf{p}}^{\kappa(m)}(z) = Q_{[\![\Sigma_\mathsf{p}^{-1}}^{\kappa(m)}(z) = \sum_{\Delta\sigma\in\Delta\backslash\Sigma_\mathsf{p}} Q_{[\![\sigma^{-1}\Delta}^{\kappa(m)}(z), \tag{5.3.8}$$

and we may attempt to write each summand $Q_{[\![\sigma^{-1}\Delta}^{\kappa(m)}(z)$ in terms of the modified Rademacher sums $Q_{\Delta,\mathsf{r}}^{\kappa(n)}(z)$ associated to Δ. Let $\Delta\sigma\in\Delta\backslash\Sigma_\mathsf{p}$ and set $\mathsf{r} = \Delta\sigma\cdot\infty$. Supposing that $w_\Gamma(\mathsf{r}) = n$,

so that $(\sigma^{-1}\Delta\sigma)_\infty = \langle T^n\rangle$, we set $\Sigma_{\mathsf{r}} = \Delta\sigma[n]$. Then we have

$$Q^{\kappa(m)}_{[\![\sigma^{-1}\Delta}(z) = Q^{\kappa(m)}_{[\![[n][1/n]\sigma^{-1}\Delta}(z) = Q^{\kappa(mn)}_{[\![\Sigma_{\mathsf{r}}^{-1}}(z), \tag{5.3.9}$$

since $Q^{\kappa(m)}_{[\![[n]\chi}(z,s) = Q^{\kappa(mn)}_{[\![\chi}(z,s)$ (cf. §4.1). Under the assumption that Δ has width one at infinity, so that $\Sigma_{\Delta\cdot\infty} = \Delta$, we may rewrite the right most term in (5.3.9) as $Q^{\kappa(mn)}_{\Delta,\mathsf{r}}(z)$. We see then that a coset $\Delta\sigma$ of Δ in Σ_{p} contributes a term $Q^{\kappa(mw_\Gamma(\mathsf{r}))}_{\Delta,\mathsf{r}}(z)$ to the right hand side of (5.3.7), where $\mathsf{r} = \Delta\sigma\cdot\infty$. According to the proof of Lemma 5.3.3 there are exactly $w_\Gamma(\mathsf{r})$ cosets $\Delta\sigma$ in $\Delta\backslash\Sigma_{\mathsf{p}}$ satisfying $\mathsf{r} = \Delta\sigma\cdot\infty$. The required identity (5.3.7) now follows. □

From the proof of Theorem 5.3.4 we see that the existence of the scaling cosets of the conclusion is verified constructively: if $\mathsf{r} = \Delta\sigma\cdot\infty$ for some coset $\Delta\sigma \in \Delta\backslash\Sigma_{\mathsf{p}}$, then we may take $\Sigma_{\mathsf{r}} = \Delta\sigma[w_\Gamma(\mathsf{r})]$.

By utilizing the inner products $\langle\cdot,\cdot\rangle^\kappa_\Delta$ of §5.2 we can reformulate Theorem 5.3.4 in a way that is independent of scaling coset choices.

Theorem 5.3.5. *Let Δ and Γ be groups commensurable with $G(\mathbb{Z})$. Suppose that Δ is a subgroup of Γ, and suppose that Δ and Γ both have width one at infinity. Let p be a cusp of Γ, let $\kappa\, min\mathbb{Z}$ such that $\kappa \le 0$ and $m < 0$. Then we have*

$$\left|\left\langle Q^{\kappa(m)}_{\Gamma,\mathsf{p}}, Q^{\kappa(n)}_{\Delta,\mathsf{r}}\right\rangle^\kappa_\Delta\right|^2 = w_\Gamma(\mathsf{r})\delta_{\Gamma,\mathsf{p}|\Gamma\mathsf{r}}\delta_{mw_\Gamma(\mathsf{r}),n} \tag{5.3.10}$$

for all $\mathsf{r} \in \mathsf{P}_\Delta$ and $n < 0$, for any choice of scaling coset systems for Δ and Γ.

To conclude this section we comment on the branching of Rademacher sums in the case that Γ both contains and normalizes Δ. Suppose then that Δ and Γ have width one at infinity, and consider the case that p is the infinite cusp $\Gamma\cdot\infty$ in (5.3.2). Then we are speaking of the map

$$\begin{aligned}\Delta\backslash\Gamma &\to \Delta\backslash\Gamma\cdot\infty\\ \Delta\gamma &\mapsto \Delta\gamma\cdot\infty,\end{aligned} \tag{5.3.11}$$

and this map is injective (and thus bijective) when Γ normalizes Δ. Indeed, since Δ is supposed to have width one at infinity we have $(\gamma^{-1}\Delta\gamma)_\infty = \Delta_\infty = B(\mathbb{Z})$, so that every coset $\Delta\gamma \in \Delta\backslash\Gamma$ is indeed a scaling coset for Δ at $\mathsf{r} = \Delta\gamma\cdot\infty$, and the injectivity of (5.3.11) follows from Lemma 5.3.1. Applying Theorem 5.3.4 now with $\mathsf{p} = \Gamma\cdot\infty$ we obtain the following result.

Proposition 5.3.6. *Suppose that Δ and Γ are groups commensurable with $G(\mathbb{Z})$ that both have width one at infinity, and suppose that Γ contains and normalizes Δ. Let $\kappa, m \in \mathbb{Z}$ such that*

$\kappa \leq 0$ and $m < 0$. Then we have

$$Q_{\Gamma}^{\kappa(m)}(z) = \sum_{\mathsf{r} \in \Delta\backslash\mathsf{p}} Q_{\Delta,\mathsf{r}}^{\kappa(m)}(z) \tag{5.3.12}$$

when the scaling cosets for Δ at the cusps $\mathsf{r} \in \Delta\backslash\Gamma\cdot\infty$ are taken to lie in $\Delta\backslash\Gamma$.

Theorem 5.3.5 now implies the following reformulation of Proposition 5.3.6 which is independent of scaling coset choices.

Proposition 5.3.7. *Suppose that Δ and Γ are groups commensurable with $G(\mathbb{Z})$ that both have width one at infinity, and suppose that Γ contains and normalizes Δ. Then we have*

$$\left|\left\langle Q_{\Gamma}^{\kappa(-1)}, Q_{\Delta,\mathsf{r}}^{\kappa(-1)}\right\rangle_{\Delta}^{\kappa}\right|^2 = \delta_{\Gamma,\mathsf{p}|\Gamma\mathsf{r}} \tag{5.3.13}$$

for all $\mathsf{r} \in \mathsf{P}_{\Delta}$, for any choice of scaling coset system for Δ.

5.4 Fractional orders

Let Γ be a group commensurable with $G(\mathbb{Z})$, let $\phi, \psi \in G(\mathbb{Q})$ and set $U = [\![\phi^{-1}\Gamma\psi$. Let us consider the problem of writing the modified Rademacher sum $Q_U^{\kappa(m)}(z)$ in terms of Rademacher sums of the form $Q_{\Gamma',\mathsf{p}'}^{\kappa(m')}(z)$ for some group Γ' and some cusp $\mathsf{p}' \in \mathsf{P}_{\Gamma'}$. According to §4.1 we have $Q_{\Gamma,\mathsf{p}|\mathsf{q}}^{\kappa(m)}(z) = Q_{\Gamma',\mathsf{p}'}^{\kappa(m)}(z)$ for $\Gamma' = \Gamma^{\mathsf{q}}$ and $\mathsf{p}' = \mathsf{p}^{\mathsf{q}}$ (cf. (4.1.9)), so we are done if we can write $Q_U^{\kappa(m)}(z)$ in terms of $Q_{\Gamma,\mathsf{p}|\mathsf{q}}^{\kappa(n)}(z)$ for some cusps $\mathsf{p},\mathsf{q} \in \mathsf{P}_{\Gamma}$ for Γ. Set $\mathsf{p} = \Gamma\phi\cdot\infty$ and $\mathsf{q} = \Gamma\psi\cdot\infty$. By the discussion of §2.6 there exist unique $\mu,\nu \in \mathbb{Q}^+$ such that the cosets $\Sigma_{\mathsf{p}} = \Gamma\phi[\mu]$ and $\Sigma_{\mathsf{q}} = \Gamma\psi[\nu]$ (cf. (2.3.3)) are scaling cosets for Γ at p and q, respectively. We then have $U = [\![[\mu]\Sigma_{\mathsf{p}}^{-1}\Sigma_{\mathsf{q}}[1/\nu]$. Let $\chi \in \Sigma_{\mathsf{q}}^{-1}\Sigma_{\mathsf{q}}$ and consider the contribution $Q_{[\![[\mu]\chi[1/\nu]}^{\kappa(m)}(z,s)$ of the coset $[\![[\mu]\chi[1/\nu] \in U$ to the modified continued Rademacher sum $Q_U^{\kappa(m)}(z,s)$. As the following lemma demonstrates, the factor $[1/\nu]$ induces a re-scaling of the input variable z.

Lemma 5.4.1. *Let $\chi \in G(\mathbb{Q})$ and let $\nu \in \mathbb{Q}^+$. Let $\kappa, m \in \mathbb{Z}$ such that $\kappa \leq 0$ and $m < 0$. Then we have*

$$Q_{[\![\chi[1/\nu]}^{\kappa(m)}(z,s) = \frac{1}{\nu^{\kappa}} Q_{[\![\chi}^{\kappa(m)}\left(\frac{z}{\nu}, s\right). \tag{5.4.1}$$

Proof. We recall from (4.1.4) that the modified continued Rademacher component function $Q_{[\![\chi[1/\nu]}^{\kappa(m)}(z,s)$ is, by definition, the difference $T_{[\![\chi[1/\nu]}^{\kappa(m)}(z,s) - T_{0,[\![\chi[1/\nu]}^{\kappa(m)}(s)$, where the functions $T_{[\![\chi[1/\nu]}^{\kappa(m)}(z,s)$ and $T_{0,[\![\chi[1/\nu]}^{\kappa(m)}(s)$ are defined by (4.1.1) and (4.1.3), respectively. We compute $\mathrm{jac}([\![\chi[1/\nu]\,, z) = \mathrm{jac}([\![\chi\,, z/\nu)/\nu$, and use the fact that $[1/\nu]$ fixes both ∞ and 0 to verify that

$$T_{[\![\chi[1/\nu]}^{\kappa(m)}(z,s) = \frac{1}{\nu^{\kappa}} T_{[\![\chi}^{\kappa(m)}\left(\frac{z}{\nu}, s\right), \quad T_{0,[\![\chi[1/\nu]}^{\kappa(m)}(s) = \frac{1}{\nu^{\kappa}} T_{0,[\![\chi}^{\kappa(m)}(s). \tag{5.4.2}$$

The claim follows from these identities. □

We now seek to describe the effect of the factor $[\mu]$ in $Q^{\kappa(m)}_{[\![[\mu]\chi}(z,s)$, for arbitrary $\chi \in G(\mathbb{Q})$. Inspecting (4.1.1) we find that

$$T^{\kappa(m)}_{[\![[\mu]\chi}(z,s) = \mu^{\kappa}\, \mathrm{e}(m\mu\chi\cdot z)\, \mathrm{t}^{\kappa}(m\mu,\chi,z,s)\, \mathrm{jac}(\chi,z)^{\kappa}, \tag{5.4.3}$$

which suggests that we generalize the notion of Rademacher sum so as to allow for fractional orders. Recall from (2.3.5) that $B(\alpha\mathbb{Z})$ denotes the subgroup of $B_u(\mathbb{Q})$ generated by T^{α} (cf. (2.3.4)). For general $\mu \in \mathbb{Q}^+$ the right hand side of (5.4.3) will not be invariant under the replacement of χ by $T\chi$, so a Rademacher sum of fractional order is not naturally defined by collections of cosets of $B(\mathbb{Z})$, but rather, by cosets of $B(h\mathbb{Z})$ for a suitably chosen positive integer $h \in \mathbb{Z}^+$.

Given $h \in \mathbb{Z}^+$ and $\chi \in G(\mathbb{Q})$ let us write $[\![^h\chi$ as a shorthand for $B(h\mathbb{Z})\chi$, and for $X \subset G(\mathbb{Q})$ let us write $[\![^h X$ as a shorthand for the set of right cosets of $B(h\mathbb{Z})$ determined by elements of X, so that

$$[\![^h X = \{B(h\mathbb{Z})\chi \mid \chi \in X\}\,. \tag{5.4.4}$$

Let $\kappa \in \mathbb{Z}$ such that $\kappa \leq 0$. We define modified and normalized fractional Rademacher sums as follows, in analogy with the constructions of §4.1. For $\mu = -g/h$ with $(g,h) = 1$ and $g, h \in \mathbb{Z}^+$, and for $[\![^h\chi \in [\![^h G(\mathbb{Q})$, we define the *continued fractional Rademacher component function of weight* 2κ *and order* μ *associated to* $[\![^h\chi$, denoted $(z,s) \mapsto T^{\kappa(\mu)}_{[\![^h\chi}(z,s)$, by setting

$$T^{\kappa(\mu)}_{[\![^h\chi}(z,s) = \mathrm{e}(\mu\, [\![^h\chi\cdot z)\, \mathrm{t}^{\kappa}(\mu, [\![^h\chi\,, z, s)\, \mathrm{jac}([\![^h\chi\,, z)^{\kappa}, \tag{5.4.5}$$

where $\mathrm{jac}([\![^h\chi\,, z) = \mathrm{jac}(\chi, z)$, and $\mathrm{t}^{\kappa}(\mu, [\![^h\chi\,, z, s)$ is the *continued fractional Rademacher regularization factor of weight* 2κ, which is in turn given by

$$\mathrm{t}^{\kappa}(\mu, [\![^h\chi\,, z, s) = \Phi(s-2\kappa, 1+s-2\kappa, \mu\, [\![^h\chi\cdot\infty - \mu\, [\![^h\chi\cdot z)(2\pi\mathbf{i}(\mu\, [\![^h\chi\cdot z - \mu\, [\![^h\chi\cdot\infty))^{s-2\kappa} \tag{5.4.6}$$

in case $[\![^h\chi \in [\![^h G(\mathbb{Q})^{\times}$, and $\mathrm{t}^{\kappa}(\mu, [\![^h\chi\,, z, s) = 1$ otherwise. We define a function $s \mapsto T^{\kappa(\mu)}_{0,[\![^h\chi}(s)$ by setting

$$\begin{aligned} T^{\kappa(\mu)}_{0,[\![^h\chi}(s) =& \mathrm{e}(\mu\, [\![^h\chi\cdot\infty)\, \big(2\pi\mathbf{i}(\mu\, [\![^h\chi\cdot 0 - \mu\, [\![^h\chi\cdot\infty)\big)^{(s-2\kappa)}\, \mathrm{jac}([\![^h\chi\,, 0)^{\kappa} \\ &- \mathrm{e}(s/2)\, \mathrm{e}(\mu\, [\![^h\chi\cdot\infty)\, \big(2\pi\mathbf{i}(\mu\, [\![^h\chi\cdot\infty - \mu\, [\![^h\chi\cdot 0)\big)^{(s-2\kappa)}\, \mathrm{jac}([\![^h\chi\,, 0)^{\kappa} \end{aligned} \tag{5.4.7}$$

in case $[\![^h\chi \in [\![G(\mathbb{Q})^{\times\times}$, and by setting $T^{\kappa(\mu)}_{0,[\![^h\chi}(s) = 0$ otherwise, and we define the *modified*

continued fractional Rademacher component function of weight 2κ *and order* μ *associated to* $[\![^h\chi$, denoted $(z,s) \mapsto Q^{\kappa(\mu)}_{[\![^h\chi}(z,s)$, by subtracting $T^{\kappa(\mu)}_{0,[\![^h\chi}(z,s)$ from $T^{\kappa(\mu)}_{[\![^h\chi}(z,s)$.

$$Q^{\kappa(\mu)}_{[\![^h\chi}(z,s) = T^{\kappa(\mu)}_{[\![^h\chi}(z,s) - T^{\kappa(\mu)}_{0,[\![^h\chi}(z,s) \tag{5.4.8}$$

For $U \subset [\![^h G(\mathbb{Q})$ we now define the *continued fractional Rademacher sum of weight* 2κ *and order* μ *associated to* U, denoted $T^{\kappa(\mu)}_U(z,s)$, and the *modified continued fractional Rademacher sum of weight* 2κ *and order* μ *associated to* U, denoted $Q^{\kappa(\mu)}_U(z,s)$, by setting

$$T^{\kappa(\mu)}_U(z,s) = \sum_{[\![^h\chi \in U} T^{\kappa(\mu)}_{[\![^h\chi}(z,s), \tag{5.4.9}$$

$$Q^{\kappa(\mu)}_U(z,s) = \sum_{[\![^h\chi \in U} Q^{\kappa(\mu)}_{[\![^h\chi}(z,s), \tag{5.4.10}$$

when these sums are absolutely locally uniformly convergent, and given such circumstances we define $T^{\kappa(\mu)}_U(z)$ and $Q^{\kappa(\mu)}_U(z)$ by taking the limit as s tends to 1 in $T^{\kappa(\mu)}_U(z,s)$ and $Q^{\kappa(\mu)}_U(z,s)$, respectively, so long as these limits exist.

$$T^{\kappa(\mu)}_U(z) = \lim_{s\to 1^+} T^{\kappa(\mu)}_U(z,s) \tag{5.4.11}$$

$$Q^{\kappa(\mu)}_U(z) = \lim_{s\to 1^+} Q^{\kappa(\mu)}_U(z,s) \tag{5.4.12}$$

We call $T^{\kappa(\mu)}_U(z)$ the *normalized fractional Rademacher sum of weight* 2κ *and order* μ *associated to* U, and we call $Q^{\kappa(\mu)}_U(z)$ the *modified fractional Rademacher sum of weight* 2κ *and order* μ *associated to* U.

We have the following vanishing result for fractional Rademacher sums.

Proposition 5.4.2. *Let* X *be a union of cosets of* $B(\mathbb{Z})$ *in* $G(\mathbb{Q})$*, let* $\mu = -g/h$ *for some* $g, h \in \mathbb{Z}^+$ *and* $(g,h) = 1$*, and set* $U = [\![^h X$*. Suppose that the normalized fractional Rademacher sum* $T^{\kappa(\mu)}_U(z)$ *and the modified fractional Rademacher sum* $Q^{\kappa(\mu)}_U(z)$ *converge. Then they vanish identically unless* $h = 1$*.*

Proof. Suppose that $h \neq 1$. By hypothesis we have a disjoint decomposition $X = \bigcup_i [\![\chi_i$ for some $\{\chi_i\} \subset X$. This implies a disjoint decomposition $X = \bigcup_i \bigcup_{k=0}^{h-1} [\![^h T^k \chi_i$, so that we have

$$\begin{aligned} T^{\kappa(\mu)}_U(z,s) &= \sum_i \sum_{k=0}^{h-1} \mathrm{e}\left(-\frac{g}{h} T^k \chi_i \cdot z\right) \mathrm{t}^\kappa\left(-\frac{g}{h}, T^k\chi_i, z, s\right) \mathrm{jac}(T^k\chi_i, z)^\kappa \\ &= \sum_i \left(\sum_{k=0}^{h-1} \mathrm{e}\left(-\frac{gk}{h}\right)\right) \mathrm{e}\left(-\frac{g}{h}\chi_i \cdot z\right) \mathrm{t}^\kappa\left(-\frac{g}{h}, \chi_i, z, s\right) \mathrm{jac}(\chi_i, z)^\kappa \end{aligned} \tag{5.4.13}$$

since the continued fractional regularization factor is unaffected when the second argument is multiplied by an element of $B(\mathbb{Z})$ on the left. The sum $\sum_{k=0}^{h-1} \mathrm{e}(-gk/h)$ vanishes for g coprime to h unless $h = 1$. This shows that the continued fractional Rademacher sum $T_U^{\kappa(\mu)}(z,s)$ vanishes for all z and s when $h \neq 1$. A directly analogous computation shows that the function $T_{0,U}^{\kappa(\mu)}(s)$ vanishes in case $h \neq 1$. We conclude that both the continued fractional Rademacher sum $T_U^{\kappa(\mu)}(z,s)$ and the modified continued fractional Rademacher sum $Q_U^{\kappa(\mu)}(z,s)$ vanish identically in case $h \neq 1$. The claim of the proposition follows. □

We return now to the identification of the Rademacher sum $Q_U^{\kappa(m)}(z)$, in the case that $U = \left[\!\left[\,[\mu]\Sigma_{\mathsf{p}}^{-1}\Sigma_{\mathsf{q}}[1/\nu]\,\right]\!\right]$.

Theorem 5.4.3. *Let Γ be a group commensurable with $G(\mathbb{Z})$, let $\mathsf{p},\mathsf{q} \in \mathsf{P}_\Gamma$ be cusps for Γ, and let Σ_{p} and Σ_{q} be scaling cosets for Γ at p and q, respectively. Let $\mu,\nu \in \mathbb{Q}^+$ and set $U = \left[\!\left[\,[\mu]\Sigma_{\mathsf{p}}^{-1}\Sigma_{\mathsf{q}}[1/\nu]\,\right]\!\right]$. Let $\kappa, m \in \mathbb{Z}$ such that $\kappa \leq 0$ and $m < 0$. Let h be the smallest positive integer such that $h\mu \in \mathbb{Z}$. If h divides m then we have*

$$Q_U^{\kappa(m)}(z) = h\frac{\mu^\kappa}{\nu^\kappa} Q_{\Gamma,\mathsf{p}|\mathsf{q}}^{\kappa(m\mu)}\left(\frac{z}{\nu}\right), \tag{5.4.14}$$

and if h does not divide m then $Q_U^{\kappa(m)}(z)$ vanishes identically.

Proof. Set $U' = \left[\!\left[\,[\mu]\Sigma_{\mathsf{p}}^{-1}\Sigma_{\mathsf{q}}\right]\!\right]$ and $U'' = \left[\!\left[\,\Sigma_{\mathsf{p}}^{-1}\Sigma_{\mathsf{q}}\right]\!\right]$, so that $U = U'[1/\nu] = [\mu]U''[1/\nu]$. By Lemma 5.4.1 we have $Q_U^{\kappa(m)}(z) = Q_{U'}^{\kappa(m)}(z/\nu)/\nu^\kappa$, so it suffices for us to show that $Q_{U'}^{\kappa(m)}(z)$ is $h\mu^\kappa Q_{\Gamma,\mathsf{p}|\mathsf{q}}^{\kappa(m\mu)}(z)$ or vanishing, according as h divides m or not.

Set $X = \Sigma_{\mathsf{p}}^{-1}\Sigma_{\mathsf{q}} \subset G(\mathbb{Q})$. Then X is a union of right cosets of $B(\mathbb{Z})$, so we have a disjoint decomposition $X = \bigcup_i \left[\!\left[\,\chi_i\right]\!\right]$ for some $\chi_i \in G(\mathbb{Q})$. By the choice of h we have $\mu = g/h$ for some $g \in \mathbb{Z}^+$ with $(g,h) = 1$. Then for the set $[\mu]X$ we have

$$[\mu]X = \bigcup_i [\mu]\left[\!\left[\,\chi_i\right]\!\right] = \bigcup_i \bigcup_{k=0}^{h-1} [\mu]T^k \left[\!\left[\,{}^h\chi_i\right]\!\right] = \bigcup_i \bigcup_{k=0}^{h-1} T^{k\mu}\left[\!\left[\,{}^g[\mu]\chi_i\right]\!\right] \tag{5.4.15}$$

with all the unions disjoint, from which we conclude that $\left[\!\left[\,[\mu]X\right]\!\right]$ admits the disjoint decomposition $\left[\!\left[\,[\mu]X\right]\!\right] = \bigcup_{k=0}^{h-1}\bigcup_i T^{k\mu}\left[\!\left[\,[\mu]\chi_i\right]\!\right]$. We have

$$Q_{T^{k\mu}\left[\!\left[\,[\mu]\chi_i\right]\!\right]}^{\kappa(m)}(z,s) = \mathrm{e}(mk\mu) Q_{\left[\!\left[\,[\mu]\chi_i\right]\!\right]}^{\kappa(m)}(z,s). \tag{5.4.16}$$

If h divides m then $m\mu \in \mathbb{Z}$ and $\mathrm{e}(mk\mu) = 1$ for all k. Further, $Q_{\left[\!\left[\,[\mu]\chi_i\right]\!\right]}^{\kappa(m)}(z,s)$ coincides with $\mu^\kappa Q_{\left[\!\left[\,\chi_i\right]\!\right]}^{\kappa(m\mu)}(z,s)$ in this case, and so we deduce the required identity $Q_{U'}^{\kappa(m)}(z) = h\mu^\kappa Q_{\Gamma,\mathsf{p}|\mathsf{q}}^{\kappa(m\mu)}(z)$. On the other hand, if h does not divide m, so that $m\mu$ is not an integer, then the sum $\sum_{k=0}^{h-1}\mathrm{e}(mk\mu)$ vanishes and this implies the vanishing of the modified Rademacher sum $Q_{U'}^{\kappa(m)}(z)$. The proof is complete. □

Taking $\mathfrak{q} = \Gamma \cdot \infty$ in Theorem 5.4.3 we obtain the following result.

Theorem 5.4.4. *Let Γ be a group commensurable with $G(\mathbb{Z})$, let Z be a finite union of left cosets of Γ in $G(\mathbb{Q})$ and set $U = [\![Z$. Let $\kappa, m \in \mathbb{Z}$ such that $\kappa \leq 0$ and $m < 0$. Then the modified Rademacher sum $Q_U^{\kappa(m)}(z)$ is an automorphic integral of weight 2κ for Γ.*

5.5 Hecke operators

For the modified Rademacher sums, in contrast to the classical Rademacher sums, it is natural to consider not only their variance with respect to group actions, but also how they vary with respect to Hecke operators. The discrepancy in constant terms between the classical and modified Rademacher sums illustrated by Proposition 4.4.10 is a stubborn barrier to an exposition of the interaction between Hecke operators and the classical Rademacher sums.

Recall that $I_\kappa(\Gamma)$ denotes the space of automorphic integrals of weight 2κ for Γ, in the sense of §2.7, and recall the operator $I_\kappa(\Gamma) \to \mathcal{O}(\mathbb{H})$ of (2.7.14), denoted $f \mapsto f\|_\Gamma^\kappa X$, and defined for a right coset $X \in \Gamma\backslash G(\mathbb{Q})$ by setting $f\|_\Gamma^\kappa X = (f - J^\kappa_{\chi\cdot\infty} g)|^\kappa \chi$ for any representative $\chi \in X$ where $g \in S_{1-\kappa}(\Gamma)$ is the cusp form associated to f (cf. (2.7.13)). Let us generalize this operator by setting

$$f\|_\Gamma^\kappa X = \sum_i (f - J^\kappa_{\chi\cdot\infty} g)|^\kappa \chi_i \tag{5.5.1}$$

in case X is a finite union of right cosets of Γ in $G(\mathbb{Q})$ and the χ_i furnish a transversal $X = \bigcup_i \Gamma\chi_i$ for X over Γ. The next lemma verifies that the operator $f \mapsto f\|_\Gamma^\kappa X$ makes sense in case X is a double coset $X = \Gamma\sigma\Gamma$ of Γ in $G(\mathbb{Q})$.

Lemma 5.5.1. *Let Γ be a group commensurable with $G(\mathbb{Z})$, let $\sigma \in G(\mathbb{Q})$, set $\Delta = \Gamma \cap \sigma^{-1}\Gamma\sigma$ and $\Delta' = \sigma\Gamma\sigma^{-1} \cap \Gamma$, and suppose that*

$$\Gamma = \bigcup_i \lambda_i' \Delta', \quad \Gamma = \bigcup_i \Delta \rho_i, \tag{5.5.2}$$

are left and right transversals for Γ over Δ' and Δ, respectively. Then the double coset $\Gamma\sigma\Gamma$ admits disjoint decompositions

$$\Gamma\sigma\Gamma = \bigcup_i \lambda_i' \sigma\Gamma, \quad \Gamma\sigma\Gamma = \bigcup_i \Gamma\sigma\rho_i, \tag{5.5.3}$$

into left and right cosets for Γ.

Proof. Since $G(\mathbb{Q})$ is the commensurator of $G(\mathbb{Z})$, the intersection $\Delta = \Gamma \cap \sigma^{-1}\Gamma\sigma$ has finite index in both Γ and $\sigma^{-1}\Gamma\sigma$, so there are only finitely many λ_i' and ρ_i. Observe next that

$\sigma\Delta = \Delta'\sigma = \sigma\Gamma \cap \Gamma\sigma$. Consequently, we have $\Delta'\sigma \subset \sigma\Gamma$ and $\sigma\Delta \subset \Gamma\sigma$, so that

$$\Gamma\sigma\Gamma = \bigcup \lambda_i'\Delta'\sigma\Gamma \subset \bigcup \lambda_i'\sigma\Gamma\Gamma = \bigcup \lambda_i'\sigma\Gamma, \tag{5.5.4}$$

$$\Gamma\sigma\Gamma = \bigcup \Gamma\sigma\Delta\rho_i \subset \bigcup \Gamma\Gamma\sigma\rho_i = \bigcup \Gamma\sigma\rho_i. \tag{5.5.5}$$

The reverse inclusions hold since all the λ_i' and ρ_i lie in Γ. The unions are disjoint, for if $\Gamma\sigma\rho_i = \Gamma\sigma\rho_j$ say, then $\rho_i\rho_j^{-1} \in \sigma^{-1}\Gamma\sigma$, but $\rho_i\rho_j^{-1} \in \Gamma$ by our choice of the ρ_i, so $\rho_i\rho_j^{-1} \in \Delta$, and this implies $i = j$. A similar argument applies to the cosets $\lambda_i'\sigma\Gamma$. □

The right hand identity of (5.5.3) shows that the operator $f \mapsto f\|_\Gamma^\kappa\Gamma\sigma\Gamma$ is well-defined for $f \in I_\kappa(\Gamma)$ and $\sigma \in G(\mathbb{Q})$. The left hand identity of (5.5.3) shows that the function $f\|_\Gamma^\kappa\Gamma\sigma\Gamma$ again lies in $I_\kappa(\Gamma)$. Indeed, if the cusp form associated to $f \in I_\kappa(\Gamma)$ is $g \in S_{1-\kappa}(\Gamma)$ then the cusp form associated to $f\|_\Gamma^\kappa\Gamma\sigma\Gamma$ is $g\|_\Gamma^\kappa\Gamma\sigma\Gamma$. We call the operator $I_\kappa(\Gamma) \to I_\kappa(\Gamma)$ given by $f \mapsto f\|_\Gamma^\kappa\Gamma\sigma\Gamma$ the *weight* 2κ *Hecke operator associated to* σ. In order to ease notation we set

$$f\dagger_\Gamma^\kappa\sigma = f\|_\Gamma^\kappa\Gamma\sigma\Gamma \tag{5.5.6}$$

for $\sigma \in G(\mathbb{Q})$. We have $f\dagger_\Gamma^\kappa\sigma = \sum_i f|^\kappa\chi_i$ when $\{\chi_i\} \subset G(\mathbb{Q})$ is a right transversal $\Gamma\sigma\Gamma = \bigcup_i \Gamma\chi_i$ for Γ in $\Gamma\sigma\Gamma$. If σ belongs to the normalizer of Γ then we have $\Gamma\sigma\Gamma = \Gamma\sigma$ and hence $f\dagger_\Gamma^\kappa\sigma = (f - J_{\sigma\cdot\infty}^\kappa g)|^\kappa\sigma$ for any $f \in I_\kappa(\Gamma)$. More generally, Lemma 5.5.1 yields for us the formula

$$f\dagger_\Gamma^\kappa\sigma = \sum_i (f - J_{\sigma\rho_i\cdot\infty}^\kappa g)|^\kappa(\sigma\rho_i) \tag{5.5.7}$$

in the case that $\{\rho_i\}$ is a right transversal for Γ over the intersection $\Delta = \Gamma \cap \Gamma^\sigma$.

The action of the operator $f \mapsto f\|_\Gamma^\kappa X$ on holomorphic Poincaré series can be described without reference to transversals. Indeed, if U is the set of right cosets of $B(\mathbb{Z})$ determined by a union of left cosets of some group Γ commensurable with $G(\mathbb{Z})$, and if X is a finite union $\bigcup_j \Gamma\chi_j$ say, of right cosets of Γ, then, taking $\kappa > 1$ to ensure absolute (and locally uniform) convergence, we have

$$P_U^{\kappa(m)}\|_\Gamma^\kappa X = \sum_j P_U^{\kappa(m)}|^\kappa\chi_j = \sum_j P_{U\chi_j}^{\kappa(m)} = P_{UX}^{\kappa(m)}, \tag{5.5.8}$$

so that the Poincaré series are stable under the action of the operators $f \mapsto f\|_\Gamma^\kappa X$. In particular, for the Hecke operator $f \mapsto f\dagger_\Gamma^\kappa\sigma$ we have $P_U^{\kappa(m)}\dagger_\Gamma^\kappa\sigma = P_{U\sigma\Gamma}^{\kappa(m)}$. This identity $P_U^{\kappa(m)}\|_\Gamma^\kappa X = P_{UX}^{\kappa(m)}$ extends naturally to the modified Rademacher sums.

Proposition 5.5.2. *Let Γ be a group commensurable with $G(\mathbb{Z})$, let Z be a finite union of left cosets of Γ in $G(\mathbb{Q})$ and set $U = [\![Z$. Let $\kappa, m \in \mathbb{Z}$ such that $\kappa \leq 0$ and $m < 0$. Then for X a*

finite union of right cosets of Γ *in* $G(\mathbb{Q})$ *we have*

$$Q_U^{\kappa(m)}\|_\Gamma^\kappa X = Q_{UX}^{\kappa(m)}. \tag{5.5.9}$$

Proof. The modified Rademacher sum $Q_U^{\kappa(m)}(z)$ is an automorphic integral of weight 2κ for Γ according to Theorem 5.4.4, so it has an associated cusp form $g \in S_{1-\kappa}(\Gamma)$ say. We may express X as a disjoint union $X = \bigcup_i \Gamma\chi_i$ for some finitely many $\chi_i \in G(\mathbb{Q})$. We then compute

$$Q_U^{\kappa(m)}\|_\Gamma^\kappa X = \sum_i (Q_U^{\kappa(m)} - J^\kappa_{\chi_i\cdot\infty} g)|^\kappa \chi_i = \sum_i Q_{U\chi_i}^{\kappa(m)} = Q_{UX}^{\kappa(m)}. \tag{5.5.10}$$

This proves the claim. □

In the case that Γ is the modular group $G(\mathbb{Z})$ we have the classical Hecke operators $\hat{T}(n)$, defined for $n \in \mathbb{Z}^+$ by setting

$$n^{1-\kappa}(\hat{T}(n)f)(z) = \sum_{\substack{ad=n\\0\le b<d}} f\left(\frac{az+b}{d}\right)\frac{a^\kappa}{d^\kappa} \tag{5.5.11}$$

for $f \in M_\kappa(\Gamma)$. We deduce the relation between the operators $\hat{T}(n)$ and $f \mapsto f\dagger_\Gamma^\kappa\sigma$, for $\Gamma = G(\mathbb{Z})$, by observing the following coincidence of disjoint unions of cosets and double cosets of Γ.

$$\bigcup_{\substack{ad=n\\0\le b<d}} \Gamma\begin{bmatrix} a & b\\ 0 & d\end{bmatrix} = \bigcup_{\substack{ad=n\\d|a}} \Gamma\begin{bmatrix} a & 0\\ 0 & d\end{bmatrix}\Gamma \tag{5.5.12}$$

Both sides of (5.5.12) are decompositions of the image in $G(\mathbb{Q})$ of the set of 2×2 matrices with integral entries and determinant n. From these decompositions we deduce the following result.

Lemma 5.5.3. *For* $n \in \mathbb{Z}^+$ *and* $\kappa \in \mathbb{Z}$ *we have*

$$n^{1-\kappa}(\hat{T}(n)f) = \sum_{d\in\mathbb{Z}^+,\, d^2|n} f\dagger_\Gamma^\kappa[n/d^2] \tag{5.5.13}$$

for any $f \in I_\kappa(\Gamma)$.

As a partial converse to Lemma 5.5.3, observe that for $\Gamma = G(\mathbb{Z})$ the action of any Hecke operator $f \mapsto f\dagger_\Gamma^\kappa\sigma$ can be expressed in terms of the Hecke operator associated to a diagonal element $[\mu] \in B(\mathbb{Q})$. Indeed, if $\sigma \in B(\mathbb{Q})$ then, according to (5.5.12), we have $\Gamma\sigma\Gamma = \Gamma[\mu]\Gamma$ for some $\mu \in \mathbb{Q}^+$. In case σ does not lie in $B(\mathbb{Q})$ we have $\sigma\cdot\infty = p \neq \infty$ for some $p \in \mathbb{Q}$. Since the modular group acts transitively on $\hat{\mathbb{Q}}$ there is some $\sigma_p \in \Gamma$ with $\sigma_p\cdot\infty = p$, and then $\Gamma\sigma\Gamma = \Gamma\tilde{\sigma}\Gamma$ where $\tilde{\sigma} = \sigma_p^{-1}\sigma$ evidently lies in $B(\mathbb{Q})$.

Lemma 5.5.4. *Let $\Gamma = G(\mathbb{Z})$ and let $\sigma \in G(\mathbb{Q})$. Then there exists $\mu \in \mathbb{Q}^+$ such that*

$$f\dagger_\Gamma^\kappa \sigma = f\dagger_\Gamma^\kappa[\mu] \tag{5.5.14}$$

for all $\kappa \in \mathbb{Z}$ and all $f \in I_\kappa(\Gamma)$.

We now consider the action of the Hecke operators $\hat{T}(n)$ on the modified Rademacher sum $Q_\Gamma^{(-1)}(z) = Q_{[\![\Gamma}^{(-1)}(z)$ of weight 0 and order -1 associated to the modular group $\Gamma = G(\mathbb{Z})$. From Theorem 4.5.3 we have that $Q_{[\![\Gamma}^{(-m)}(z)$ is an automorphic integral of weight 0 (that is, an abelian integral) for Γ for $m > 0$. Since there are no non-zero cusp forms of weight 2 for the modular group, we see from Corollary 4.5.5 that $Q_{[\![\Gamma}^{(-m)}(z)$ is in fact a Γ-invariant function on $\mathbb{H}$ for all $m \in \mathbb{Z}^+$. Consequently we have $Q_{[\![\Gamma}^{(-1)}\dagger_\Gamma^0[n] = Q_{[\![\Gamma[n]\Gamma}^{(-1)}$ for $n \in \mathbb{Z}^+$, by Proposition 5.5.2 and the definition (5.5.6) of $f\dagger_\Gamma^\kappa\sigma$. Suppose that n is square-free. Then, according to Lemma 5.5.3, the action of the operator $f \mapsto f\dagger_\Gamma^0[n]$ on $M_0(\Gamma)$ coincides with that of $n\hat{T}(n)$. We anticipate an application of Lemma 5.5.1. For $\sigma = [n]$ we have $\Delta' = \sigma\Gamma\sigma^{-1} \cap \Gamma = \Gamma^0(n)$, and for a left transversal of $\Delta' = \Gamma^0(n)$ in Γ we may take

$$\Gamma = \bigcup_{e\|n} \bigcup_{k=0}^{n/e-1} T^k S T^e \Delta' \tag{5.5.15}$$

where the first union is over exact divisors of n (cf. §2.1). We deduce that

$$\Gamma[n]\Gamma = \bigcup_{e\|n} \bigcup_{k=0}^{n/e-1} T^k S T^e [n]\Gamma \tag{5.5.16}$$

by Lemma 5.5.1. Now $[\![T^k X = [\![X$ for any subset $X \subset G(\mathbb{Q})$ and any $k \in \mathbb{Z}$, so we find that

$$n(\hat{T}(n)Q_{[\![\Gamma}^{(-1)}) = Q_{[\![\Gamma[n]\Gamma}^{(-1)} = \sum_{e\|n} Q_{[\![ST^e[n]\Gamma}^{(-1)}. \tag{5.5.17}$$

Lemma 5.5.5. *Let e be an exact divisor of n. Then we have $[\![ST^e[n]\Gamma = [\![[n/e^2]\Gamma$ when Γ is the modular group $G(\mathbb{Z})$.*

Proof. We compute

$$ST^e[n] = \begin{bmatrix} 0 & -1 \\ n & e \end{bmatrix}, \quad ST^e[n]ST^{n/e}S = \begin{bmatrix} n/e & -1 \\ 0 & e \end{bmatrix}, \tag{5.5.18}$$

and observe that $T^k A T^l = [n/e^2]$ for $k, l \in \mathbb{Z}$ such that $ke + ln/e = 1$, where A is given by $A = ST^e[n]ST^{n/e}S$. □

We can now write

$$n(\hat{T}(n)Q_{[\![\Gamma}^{(-1)}) = \sum_{e\|n} Q_{[\![[n/e^2]\Gamma}^{(-1)} \tag{5.5.19}$$

in the case that n is square-free, and we can employ the methods of §5.4 to rewrite (5.5.19) in terms of the fractional Rademacher sums. Observe that $B(g\mathbb{Z})[g/h] = [g/h]B(h\mathbb{Z})$. We thus have

$$[\![[\mu]\chi = \bigcup_{k=0}^{g-1} T^k\, [\![^g[\mu]\chi = \bigcup_{k=0}^{g-1} T^k[\mu]\, [\![^h\chi \tag{5.5.20}$$

for any $\chi \in G(\mathbb{Q})$ when $\mu = g/h$ with $g, h \in \mathbb{Z}^+$ and $(g,h) = 1$. We rewrite the right hand side of (5.5.19), taking $g = n/h$ and $h = e$, as

$$\sum_{h\|n} Q_{[\![[n/h^2]\Gamma}^{(-1)} = \sum_{h\|n} Q_{[n/h^2][\![^h\Gamma}^{(-1)} = \sum_{h\|n} Q_{[\![^h\Gamma}^{(-n/h^2)}, \tag{5.5.21}$$

and thus obtain the following result, relating the actions of Hecke operators to the fractional Rademacher sums.

Theorem 5.5.6. *Let $n \in \mathbb{Z}^+$ be square-free and let $\Gamma = G(\mathbb{Z})$. Then we have*

$$n(\hat{T}(n)Q_{[\![\Gamma}^{(-1)}) = \sum_{h\|n} Q_{[\![^h\Gamma}^{(-n/h^2)} \tag{5.5.22}$$

where the sum is over the exact divisors of n.

It is interesting to compare the result (5.5.22) of Theorem 5.5.6 with the identity

$$n\left(\hat{T}(n)Q_{[\![\Gamma}^{(-1)}\right)(z) = Q_{[\![\Gamma}^{(-n)}(z), \tag{5.5.23}$$

which holds since both sides are holomorphic on $\mathbb{H}$, and have the same singular terms in their Fourier expansion at $\Gamma\cdot\infty$. This proves that the Rademacher sum $Q_U^{(n/h^2)}(z)$ vanishes for $U = [\![^h\Gamma$ whenever h is an exact divisor of n, and thus recovers a special case of the vanishing result of Theorem 5.4.3.

6 Moonshine

Monstrous moonshine associates a group Γ_g commensurable with $G(\mathbb{Z})$ and having width one at infinity to each (conjugacy class of) element(s) g in the Monster group $\mathbb{M}$. In the article [CMS04] the set of groups $\{\Gamma_g \mid g \in \mathbb{M}\}$ is characterized in purely group theoretic terms. In this

section we describe a reformulation of this characterization in terms of normalized Rademacher sums, and a certain family of moduli problems for solid tori with conformal structure on the boundary.

6.1 Genera

The main theorem of [CMS04] gives four conditions which, taken together, characterize the groups Γ_g for $g \in \mathbb{M}$. The first of these conditions is the requirement that a group Γ have genus zero. Our first result in this section is a reformulation of the genus zero condition in terms of Rademacher sums.

Theorem 6.1.1. *Let Γ be a group commensurable with $G(\mathbb{Z})$ that has width one at infinity and let $\mathfrak{p}$ be a cusp for Γ. Then the normalized Rademacher sum $T^{(-1)}_{\Gamma,\mathfrak{p}}(z)$ defines a Γ-invariant function on $\mathbb{H}$ if and only if Γ has genus zero.*

Proof. Let Γ be as in the statement of the theorem and let $\mathfrak{p}, \mathfrak{q} \in \mathsf{P}_\Gamma$ be cusps of Γ. By Theorem 4.4.6 we have

$$T^{(-1)}_{\Gamma,\mathfrak{p}|\mathfrak{q}}(z) = \delta_{\Gamma,\mathfrak{p}|\mathfrak{q}}\, \mathrm{e}(-z) + F^{(-1)}_{\Gamma,\mathfrak{p}|\mathfrak{q}}(z)_{\mathrm{van}}. \tag{6.1.1}$$

Taking $\mathfrak{q} = \Gamma \cdot \infty$ in (6.1.1) we see that $T^{(-1)}_{\Gamma,\mathfrak{p}}(z)$ is holomorphic on $\mathbb{H}$, and letting $\mathfrak{q}$ range over P_Γ we see from (6.1.1) that the only pole of $T^{(-1)}_{\Gamma,\mathfrak{p}}(z)$ is a simple pole at the cusp $\mathfrak{p}$. We see then that if the function $T^{(-1)}_{\Gamma,\mathfrak{p}}(z)$ is Γ-invariant then it defines a morphism ϕ say, of Riemann surfaces $\phi : \mathsf{X}_\Gamma \to \hat{\mathbb{C}}$ (cf. §2.5) which has degree one since the preimage of $\infty \in \hat{\mathbb{C}}$ under ϕ is the single point $\mathfrak{p} \in \mathsf{X}_\Gamma$ and ϕ is unramified here since Γ is assumed to have width one at infinity. We conclude that ϕ is an isomorphism, so that Γ indeed has genus zero.

Conversely, if Γ has genus zero then, since the space $S_1(\Gamma)$ of cusp forms of weight 2 for Γ is isomorphic to the space of holomorphic differentials on X_Γ, we have $S_1(\Gamma) = \{0\}$, so that any automorphic integral of weight 0 for Γ is in fact a Γ-invariant function, by Corollary 4.5.5. The modified Rademacher sum $Q^{(-1)}_{\Gamma,\mathfrak{p}}(z)$ is an automorphic integral for Γ by Theorem 4.5.3. The normalized Rademacher sum $T^{(-1)}_{\Gamma,\mathfrak{p}}(z)$ differs from $Q^{(-1)}_{\Gamma,\mathfrak{p}}(z)$ by a constant function by Proposition 4.4.11, and thus $T^{(-1)}_{\Gamma,\mathfrak{p}}(z)$ is Γ-invariant whenever $Q^{(-1)}_{\Gamma,\mathfrak{p}}(z)$ is. We conclude that $T^{(-1)}_{\Gamma,\mathfrak{p}}(z)$ is Γ-invariant in case Γ has genus zero. This completes the proof. □

From the first part of the proof of Theorem 6.1.1 we see that the normalized Rademacher sum $T^{(-1)}_{\Gamma,\mathfrak{p}}(z)$ associated to Γ at a cusp $\mathfrak{p}$ defines an isomorphism $\mathsf{X}_\Gamma \to \hat{\mathbb{C}}$ mapping $\mathfrak{p}$ to $\infty \in \hat{\mathbb{C}}$ in case Γ is a group of genus zero.

Theorem 6.1.2. *Let Γ be a group commensurable with $G(\mathbb{Z})$ that has width one at infinity and let $\mathfrak{p} \in \mathsf{P}_\Gamma$ be a cusp of Γ. If Γ has genus zero then the normalized Rademacher sum $T^{(-1)}_{\Gamma,\mathfrak{p}}(z)$ associated to Γ at the cusp $\mathfrak{p}$ induces an isomorphism $\mathsf{X}_\Gamma \to \hat{\mathbb{C}}$ mapping $\mathfrak{p}$ to ∞.*

Applying Theorem 6.1.2 with $\mathsf{p} = \Gamma \cdot \infty$ the infinite cusp we see that the normalized Rademacher sum $T_\Gamma^{(-1)}(z)$ induces an isomorphism $\mathsf{X}_\Gamma \to \hat{\mathbb{C}}$ mapping the infinite cusp to $\infty \in \hat{\mathbb{C}}$. Applying Proposition 4.4.9 with $\mathsf{p} = \mathsf{q} = \Gamma \cdot \infty$ we see that the Fourier expansion of $T_\Gamma^{(-1)}(z)$ has vanishing constant term. This shows that the normalized Rademacher sum $T_\Gamma^{(-1)}(z)$ associated to Γ at the infinite cusp is the normalized hauptmodul for Γ (cf. §2.5) whenever this statement makes sense; i.e. whenever Γ has genus zero.

Theorem 6.1.3. *Let Γ be a group commensurable with $G(\mathbb{Z})$ that has width one at infinity. If Γ has genus zero then the normalized Rademacher sum $T_\Gamma^{(-1)}(z)$ is the normalized hauptmodul for Γ.*

From the second part of the proof of Theorem 6.1.1 we see that if Γ has genus zero then not only the first order but in fact all the higher order normalized Rademacher sums $T_{\Gamma,\mathsf{p}}^{(-m)}(z)$ (for $m > 0$) associated to Γ at an arbitrary cusp $\mathsf{p} \in \mathsf{P}_\Gamma$ are Γ-invariant.

Theorem 6.1.4. *Let Γ be a subgroup of $G(\mathbb{Q})$ that is commensurable with $G(\mathbb{Z})$ and has width one at infinity and let $\mathsf{p} \in \mathsf{P}_\Gamma$ be a cusp of Γ. Let $m \in \mathbb{Z}^+$. If Γ has genus zero then the normalized Rademacher sum $T_{\Gamma,\mathsf{p}}^{(-m)}(z)$ is Γ-invariant.*

6.2 Moduli

The second condition of the main theorem of [CMS04] is that a group Γ be of $n\|h$-type (cf. §2.5). In this section we provide a geometric description of the groups of $n\|h$-type by furnishing a family of moduli problems for which the corresponding moduli spaces may be realized as quotients of the form $\mathsf{Y}_\Gamma = \Gamma\backslash\mathbb{H}$ where Γ is a group of $n\|h$-type.

Consider pairs (E, C) where E is an elliptic curve over $\mathbb{C}$ and C is an oriented subgroup of E isomorphic to S^1. We call such a pair a *solid torus*. For (E, C) a solid torus, we call E the *underlying elliptic curve*, and we call C the *underlying primitive cycle*. Note that for (E, C) a solid torus, the group C determines a primitive element of the first homology group of (the smooth real surface underlying) the elliptic curve E.

Given $z \in \mathbb{H}$ let us write Λ_z for the lattice $\mathbb{Z}z + \mathbb{Z} \subset \mathbb{C}$. Observe that the lattices Λ_z and $\Lambda_{z'}$ coincide if any only if $z - z' \in \mathbb{Z}$, so the assignment $z \mapsto \Lambda_z$ descends naturally to the orbit space $B(\mathbb{Z})\backslash\mathbb{H}$. If we agree to write $[\![z$ as a shorthand for the $B(\mathbb{Z})$ orbit determined by $z \in \mathbb{H}$, then we may unambiguously write $\Lambda_{[\![z}$ for the lattice $\mathbb{Z}z + \mathbb{Z}$. We write $E_{[\![z}$ for the corresponding elliptic curve $E_{[\![z} = \mathbb{C}/\Lambda_{[\![z} = \mathbb{C}/\mathbb{Z}z + \mathbb{Z}$. Then the pair $(E_{[\![z}, C_{[\![z})$ is a solid torus when we take the subgroup $C_{[\![z}$ to be the image of $\mathbb{R}$ in $E_{[\![z}$ under the composition $\mathbb{R} \to \mathbb{C} \to E_{[\![z}$. Observe that for any solid torus (E, C) there exists $z \in \mathbb{H}$ and a unique isomorphism of elliptic curves $E \to E_{[\![z}$ inducing an orientation preserving isomorphism $C \to C_{[\![z}$.

For (E, C) a solid torus and $n \in \mathbb{Z}^+$ we write $E[n]$ for the group of n-division points of E, and we write $C[n]$ for the intersection $C \cap E[n]$. Since C is oriented, each group $C[n]$ comes

equipped with a distinguished generator; viz., the point corresponding to $1/n + \Lambda_{\mathbb{I}z}$ under an isomorphism $E \to E_{\mathbb{I}z}$ inducing an orientation preserving isomorphism $C \to C_{\mathbb{I}z}$. If K is a subgroup of $E[n]$ for some n then the quotient $\bar{E} = E/K$ is again an elliptic curve, and the image $\bar{C}$ say, of C under the natural map $E \to \bar{E}$ defines a primitive cycle of $\bar{E}$, so that the pair $(\bar{E}, \bar{C})$ is again a solid torus. We call this the *quotient of* (E, C) *by* K and denote it $(E, C)/K$.

For $n \in \mathbb{Z}^+$ define an *n-compatible isogeny of solid tori* $(E', C') \to (E, C)$ to be an isogeny $E' \to E$ of elliptic curves that maps $C'[n]$ to a subgroup of $C[n]$. Then a 1-compatible isogeny of solid tori is just an isogeny of the underlying elliptic curves. We define an *isogeny of solid tori* $(E', C') \to (E, C)$ to be an isogeny $E' \to E$ of elliptic curves that restricts to an orientation preserving map $C' \to C$ on the underlying primitive cycles. For (E, C) a solid torus and $n \in \mathbb{Z}^+$, we may now interpret the canonical map $E \to E/C[n]$ as defining an isogeny $(E, C) \to (E, C)/C[n]$ of solid tori. We call this isogeny the *canonical n-fold quotient of* (E, C).

Observe that the elliptic curve $E/C[n]$ underlying the quotient $(E, C)/C[n]$ both receives a natural map from E and maps naturally to E, for the quotient $E/E[n]$ is naturally isomorphic to E, and for the map $E/C[n] \to E$ we may take the composition $E/C[n] \to E/E[n] \xrightarrow{\sim} E$ where the first map is the natural projection and the second map is the isomorphism just mentioned. This map $E/C[n] \to E$ defines an isogeny $(E, C)/C[n] \to (E, C)$ of solid tori; indeed, it restricts to an isomorphism on the underlying primitive cycles. We call the isogeny $(E, C)/C[n] \to (E, C)$ of solid tori the *canonical n-fold cover of* (E, C).

Say an n-compatible isogeny of solid tori $(E', C') \to (E, C)$, for $n \in \mathbb{Z}^+$, is an *n-compatible isomorphism of solid tori* in case it admits an inverse n-compatible isogeny; i.e. an isogeny $E \to E'$ mapping $C[n]$ to a subgroup of $C'[n]$ such that the compositions $E' \to E \to E'$ and $E \to E' \to E$ are the identity maps on E' and E, respectively. Simply put, an n-compatible isomorphism of solid tori $(E', C') \to (E, C)$ is an isomorphism of the underlying elliptic curves $E' \to E$ that induces an isomorphism of groups $C'[n] \to C[n]$. An *isomorphism of solid tori* $(E', C') \to (E, C)$ is an isomorphism $E' \to E$ of elliptic curves that restricts to an orientation preserving isomorphism $C' \to C$ on the underlying primitive cycles.

Suppose (E, C) and (E', C') are solid tori. For n a positive integer and e an exact divisor of n, say (E, C) and (E', C') are *$n+e$-related* if there is an n-compatible isogeny $(E', C') \to (E, C)$ that induces n/e-compatible isomorphisms $(E', C')/C'[e] \to (E, C)$ and $(E', C') \to (E, C)/C[e]$. More precisely, we require that the n-compatible isogeny $(E', C') \to (E, C)$ factor through the canonical e-fold maps $(E', C') \to (E', C')/C'[e]$ and $(E, C)/C[e] \to (E, C)$, defining n/e-

compatible isomorphisms of the specified type.

$$\begin{array}{ccc} (E',C') & \xrightarrow[n/e]{\sim} & (E,C)/C[e] \\ \downarrow & & \downarrow \\ (E',C')/C'[e] & \xrightarrow[n/e]{\sim} & (E,C) \end{array} \tag{6.2.1}$$

Then solid tori (E,C) and (E',C') are $n+1$-related just if they are n-compatible isomorphic, which is the case just if there is an isomorphism of elliptic curves $E' \to E$ mapping $C'[n]$ onto $C[n]$. In particular, the notion of being $n+1$-related is an equivalence relation on solid tori. For S a subset of the set of exact divisors of n, say solid tori (E,C) and (E',C') are $n+S$-*related* if (E,C) and (E',C') are $n+e$-related for some $e \in S$.

Recall from §2.2 that the set of exact divisors of n, denoted $\mathrm{Ex}(n)$, admits a naturally defined group structure.

Lemma 6.2.1. *Let $n \in \mathbb{Z}^+$ and let $S \subset \mathrm{Ex}(n)$. Then the $n+S$-relation is an equivalence relation on solid tori if and only if S is a subgroup of* $\mathrm{Ex}(n)$.

In light of Lemma 6.2.1, we say that solid tori (E,C) and (E',C') are $n+S$-*equivalent* if they are $n+S$-related and S is a subgroup of the group of exact divisors of n.

In a slight modification of the notation of [CN79], we write $\Gamma_0(n)+S$ for the group formed by taking the union of the Hecke congruence group $\Gamma_0(n)$ and the Atkin–Lehner involutions $W_e(n)$ for e in S, when S is a subgroup of $\mathrm{Ex}(n)$.

$$W_e(n) = \left\{ \begin{bmatrix} ae & b \\ cn & de \end{bmatrix} \in G(\mathbb{Q}) \mid a,b,c,d \in \mathbb{Z},\ ade - bcn/e = 1 \right\} \tag{6.2.2}$$

Let n be a positive integer and let h be a divisor of n. Then an n-compatible isomorphism $\phi : (E',C') \to (E,C)$ naturally determines an nh-compatible isogeny $\hat{\phi} : (E',C') \to (E,C)$ which factors through an n/h-compatible isomorphism $\bar{\phi} : (E',C')/C'[h] \to (E,C)/C[h]$ on the canonical h-fold quotients, via the canonical maps $(E',C') \to (E',C')/C'[h]$ and $(E,C)/C[h] \to (E,C)$, as depicted in (6.2.3).

$$\begin{array}{ccc} (E',C') & \xrightarrow[nh]{\hat{\phi}} & (E,C) \\ & \searrow \quad \nearrow & \\ (E',C')/C'[h] & \xrightarrow[n/h]{\bar{\phi}} & (E,C)/C[h] \end{array} \tag{6.2.3}$$

We may take this nh-compatible isogeny $\hat{\phi}$ to be the composition $E' \to E \to E/E[h] \to E$ where the first map is the given map ϕ, the second map is the natural projection, and the third map is the natural isomorphism. Since ϕ is an n-compatible isomorphism the composition $E' \to E \to E/E[h]$ factors through the natural projection $E/C[h] \to E/E[h]$, and the kernel of the induced map $E' \to E/C[h]$ is just $C'[h]$, and so we arrive at an isomorphism $E'/C'[h] \to E/C[h]$ of elliptic curves, which defines the required n/h-compatible isomorphism $\bar{\phi} : (E', C')/C'[h] \to (E, C)/C[h]$ of the corresponding solid tori.

$$\begin{array}{ccccccc} (E',C') & \xrightarrow[n]{\phi} & (E,C) & \longrightarrow & (E,C)/E[h] & \xrightarrow{\sim} & (E,C) \\ & \searrow & & & \uparrow & \nearrow & \\ & (E',C')/C'[h] & \xrightarrow[n/h]{\bar{\phi}} & & (E,C)/C[h] & & \end{array} \tag{6.2.4}$$

More generally, and by the same argument, an n-compatible isogeny $(E', C') \to (E, C)$ naturally defines an nh-compatible isogeny $(E', C') \to (E, C)$ which factors through an n/h-compatible isogeny $(E', C')/C'[h] \to (E, C)/C[h]$, via the canonical maps $(E', C') \to (E', C')/C'[h]$ and $(E, C)/C[h] \to (E, C)$.

Not every n/h-compatible isogeny $\bar{\phi} : (E', C')/C'[h] \to (E, C)/C[h]$ arises in this way from an n-compatible isogeny $\phi : (E', C') \to (E, C)$, but, given an n/h-compatible isogeny $\bar{\phi} : (E', C')/C'[h] \to (E, C)/C[h]$, we always have the nh-compatible isogeny $\hat{\phi} : (E', C') \to (E, C)$ obtained as the composition

$$(E', C') \to (E', C')/C'[h] \to (E, C)/C[h] \to (E, C) \tag{6.2.5}$$

where the first map is the canonical h-fold quotient, and the third map is the canonical h-fold cover. We call this composition $\hat{\phi}$ the *nh-compatible isogeny induced from* $\bar{\phi}$.

For $n \in \mathbb{Z}^+$ and h a divisor of n say solid tori (E, C) and (E', C') are *$n|h$-related* if the canonical h-fold quotients $(E, C)/C[h]$ and $(E', C')/C'[h]$ are n/h-compatible isomorphic. For e an exact divisor of n/h, say (E, C) and (E', C') are *$n|h + e$-related* if the quotients $(E, C)/C[h]$ and $(E', C')/C'[h]$ are $n/h + e$-related. For S a subset of $\mathrm{Ex}(n/h)$ say (E, C) and (E', C') are *$n|h + S$-related* if they are $n|h + e$-related for some $e \in S$.

Lemma 6.2.2. *Let $n \in \mathbb{Z}^+$, let h be a divisor of n, and let $S \subset \mathrm{Ex}(n/h)$. Then the $n|h + S$-relation is an equivalence relation on solid tori if and only if S is a subgroup of* $\mathrm{Ex}(n/h)$.

Let N be a positive integer, set $h = ((N, 24)$ (cf. §2.2), and set $n = N/h$. Then for S a subgroup of the group of exact divisors of n/h we arrive at the notion of $n\|h + S$-equivalence of solid tori in the following way. Suppose (E, C) and (E', C') are $\Gamma_0(n|h) + S$-equivalent. Then we

have isomorphisms $(E,C) \to (E_{\mathbb{I}z}, C_{\mathbb{I}z})$ and $(E',C') \to (E_{\mathbb{I}z'}, C_{\mathbb{I}z'})$ for some $z, z' \in \mathbb{H}$. Let S' be the image of S under the natural injection $\mathrm{Ex}(n/h) \to \mathrm{Ex}(N)$ (cf. §2.2), and define elements x, y and $w_{e'}$ for $e' \in S'$ by setting

$$x = \begin{bmatrix} 1 & 1/h \\ 0 & 1 \end{bmatrix}, \quad y = \begin{bmatrix} 1 & 0 \\ n & 1 \end{bmatrix}, \quad w_{e'} = \begin{bmatrix} ae' & b \\ cN & de' \end{bmatrix}, \tag{6.2.6}$$

where the a, b, c and d in the definition of $w_{e'}$ are arbitrary integers for which we have $ade' - bcN/e' = 1$. Then there is a sequence $(\gamma_0, \gamma_1, \cdots, \gamma_k)$ with $\gamma_0 \in \Gamma_0(N)$ and $\gamma_i \in \{x, y, w_{e'} \mid e' \in S'\}$ for $1 \le i \le k$ for which we have $z' = \gamma_k \cdots \gamma_1 \gamma_0 \cdot z$. We say now that (E,C) and (E',C') are $\Gamma_0(n\|h) + S$-*equivalent* in the case that the sum $\sum_{i=1}^{k} \lambda(\gamma_i)$ vanishes in $\mathbb{Z}/h$, where $\lambda(x) = 1$, and $\lambda(w_{e'}) = 0$ for all $e' \in S'$, and $\lambda(y)$ is -1 or 1 according as $\Gamma_0(n|h) + S$ is Fricke or not.

Theorem 6.2.3. *Let $N \in \mathbb{Z}^+$, set $h = ((N, 24)$ and set $n = N/h$. Let S be a subgroup of $\mathrm{Ex}(n/h)$, and set $\Gamma = \Gamma_0(n\|h) + S$. Then the quotient $\Gamma\backslash\mathbb{H}$ is in natural correspondence with $n\|h + S$-equivalence classes of solid tori.*

6.3 Exponents

The third of the four conditions of the main theorem of [CMS04] states that if Γ is of the form $\Gamma_0(n\|h)+S$ then the quotient group $\Gamma/\Gamma_0(nh)$ should be a group of exponent two. In the present section we investigate the relationship between this exponent two condition and properties of scaling cosets.

Lemma 6.3.1. *Suppose that Δ is a group commensurable with $G(\mathbb{Z})$ that has width one at infinity and let $\mathsf{r} \in \mathsf{P}_\Delta$ be a cusp of Δ. Let $\kappa, m \in \mathbb{Z}$ such that $\kappa \le 0$ and $m < 0$. If Σ_r is a scaling coset for Δ at r with the property that $\Sigma_\mathsf{r}^2 = \Delta$ then*

$$Q_{\Delta,\mathsf{r}}^{\kappa(m)}(z) = Q_{\Delta|\mathsf{r}}^{\kappa(m)}(z), \tag{6.3.1}$$

and Σ_r is contained in the normalizer of Δ.

Proof. Set $U = [\![\Sigma_\mathsf{r}^{-1}$ and set $U' = [\![\Sigma_\mathsf{r}$. Then, under the assumption that Δ has width one at infinity, we have $Q_{\Delta,\mathsf{r}}^{\kappa(m)}(z) = Q_U^{\kappa(m)}(z)$ and $Q_{\Delta|\mathsf{r}}^{\kappa(m)}(z) = Q_{U'}^{\kappa(m)}(z)$. Let $\sigma \in \Sigma_\mathsf{r}$. Then $\Sigma_\mathsf{r}^2 = \Delta$ implies $\sigma\Delta\sigma = \Delta$, which is equivalent to the identity $\Delta\sigma = \sigma^{-1}\Delta$. We conclude that $\Sigma_\mathsf{r}^{-1} = \Sigma_\mathsf{r}$, so that $U = U'$, so that $Q_{\Delta,\mathsf{r}}^{\kappa(m)}(z) = Q_{\Delta|\mathsf{r}}^{\kappa(m)}(z)$. The identity $\sigma\Delta\sigma = \Delta$ also implies $\sigma^2 \in \Delta$, so that $\sigma\Delta\sigma = \sigma^{-1}\Delta\sigma$. We conclude that σ normalizes Δ, so that $\Sigma_\mathsf{r} = \Delta\sigma \in \Delta\backslash N(\Delta)$. This completes the proof. □

Lemma 6.3.1 states that if a scaling coset at a cusp r of a group Δ can be chosen so that its square is the trivial coset then we have a kind of commutativity for the Rademacher sums

associated to Δ; viz. the expansion at infinity of the modified Rademacher sum associated to Δ and the cusp r coincides with the expansion at r of the modified Rademacher sum associated to Δ and the cusp at infinity.

We also have the following converse to Lemma 6.3.1.

Proposition 6.3.2. *Suppose that Δ is a group commensurable with $G(\mathbb{Z})$ that has width one at infinity and let $\mathsf{r} \in \mathsf{P}_\Delta$ be a cusp of Δ. Let $\kappa, m \in \mathbb{Z}$ such that $\kappa \leq 0$ and $m < 0$. If Σ_r is a scaling coset for Δ at r that is contained in the normalizer of Δ and*

$$Q^{\kappa(m)}_{\Delta,\mathsf{r}}(z) = Q^{\kappa(m)}_{\Delta|\mathsf{r}}(z) \tag{6.3.2}$$

then $\Sigma^2_\mathsf{r} = \Delta$.

Proof. As in the proof of Lemma 6.3.1 we set $U = [\![\Sigma^{-1}_\mathsf{r}$ and $U' = [\![\Sigma_\mathsf{r}$. Then since Δ is assumed to have width one at infinity we have $Q^{\kappa(m)}_{\Delta,\mathsf{r}}(z) = Q^{\kappa(m)}_U(z)$ and $Q^{\kappa(m)}_{\Delta|\mathsf{r}}(z) = Q^{\kappa(m)}_{U'}(z)$. Let $\sigma \in \Sigma_\mathsf{r}$. Then we have $U = [\![\sigma^{-1}\Delta$ and $U' = [\![\Delta\sigma = [\![\sigma\Delta$. Set $\mathsf{r}' = \Delta\sigma^{-1} \cdot \infty$. Then $Q^{\kappa(m)}_{U'}(z) = Q^{\kappa(m)}_{\Delta,\mathsf{r}'}(z)$, so the coincidence $Q^{\kappa(m)}_U(z) = Q^{\kappa(m)}_{U'}(z)$ implies that $\mathsf{r} = \mathsf{r}'$. This in turn implies $\sigma^2 \in \Delta$, so that $\Sigma^2_\mathsf{r} = \Delta$, as we required to show. □

6.4 Cusps

The forth condition of the main theorem of [CMS04] is the following. Supposing that a group Γ is of the form $\Gamma = \Gamma_0(n\|h) + S$ for some subgroup S of the group of exact divisors of n/h, for each $p \in \hat{\mathbb{Q}}$ there should exist an element $\tilde{\sigma}_p \in G(\mathbb{R})$ such that

$$\infty = \tilde{\sigma}_p \cdot p, \quad (\tilde{\sigma}_p \Gamma \tilde{\sigma}_p^{-1})_\infty = B(\mathbb{Z}), \quad \tilde{\sigma}_p \Gamma \tilde{\sigma}_p^{-1} \supset \Gamma_0(nh). \tag{6.4.1}$$

Observe that if we set $\sigma_p = \tilde{\sigma}_p^{-1}$, then the conditions of (6.4.1) translate into

$$\sigma_p \cdot \infty = p, \quad (\sigma_p^{-1} \Gamma \sigma_p)_\infty = B(\mathbb{Z}), \quad \sigma_p^{-1} \Gamma \sigma_p \supset \Gamma_0(nh), \tag{6.4.2}$$

the first two of which are just the conditions (2.6.1) of Lemma 2.6.1 except that we allow σ_p to lie in $G(\mathbb{R})$ in (6.4.2) but insist that σ_p belong to $G(\mathbb{Q})$ in (2.6.1). Actually, an element $\sigma_p \in G(\mathbb{R})$ satisfying (6.4.2) must lie in $G(\mathbb{Q})$, as the following result demonstrates.

Lemma 6.4.1. *Let Δ and Γ be groups commensurable with $G(\mathbb{Z})$ and suppose that Δ has width one at infinity. Let $p \in \hat{\mathbb{Q}}$ and suppose that $\sigma \in G(\mathbb{R})$ satisfies $\sigma \cdot \infty = p$ and $\sigma^{-1}\Gamma\sigma \supset \Delta$. Then $\sigma \in G(\mathbb{Q})$.*

Proof. Suppose that $\sigma \in G(\mathbb{R})$ satisfies the hypotheses of the lemma. Then $(\sigma^{-1}\Gamma\sigma)_\infty$ is an infinite cyclic group containing Δ_∞ and $\Delta_\infty = B(\mathbb{Z})$ since Δ has width one at infinity. It must be

then that $(\sigma^{-1}\Gamma\sigma)_\infty$ is generated by $T^{1/n}$ for some $n \in \mathbb{Z}^+$, so that $([n]\sigma^{-1}\Gamma\sigma[1/n])_\infty = B(\mathbb{Z})$. According to Lemma 2.6.1 there exists $\sigma' \in G(\mathbb{Q})$ such that $\sigma' \cdot \infty = p$ and $((\sigma')^{-1}\Gamma\sigma')_\infty = B(\mathbb{Z})$. Now $[n]\sigma^{-1}\sigma'$ fixes ∞ and so lies in $B(\mathbb{R})$. Since σ' and $\sigma[1/n]$ both conjugate Γ to a group with width one at infinity we must in fact have $[n]\sigma^{-1}\sigma' \in B_u(\mathbb{R})$, where $B_u(\mathbb{R})$ consists of all the elements T^α (cf. (2.3.4)) with $\alpha \in \mathbb{R}$. We conclude that $\sigma = \sigma'' T^\alpha$ for some $\sigma'' \in G(\mathbb{Q})$ and $\alpha \in \mathbb{R}$. Let $\gamma \in \Gamma$ such that $\sigma^{-1}\gamma\sigma \in \Delta^\times$, so that $\sigma^{-1}\gamma\sigma \cdot \infty = q$ for some $q \in \mathbb{Q}$. Then $q = q'' - \alpha$ for $q'' = (\sigma'')^{-1}\gamma\sigma'' \cdot \infty$. Since both q and q'' lie in $\mathbb{Q}$ we conclude that α also lies in $\mathbb{Q}$. Then the identity $\sigma = \sigma'' T^\alpha$ implies that $\sigma \in G(\mathbb{Q})$, as we required to show. □

Recall from §2.6 that for Γ a group commensurable with $G(\mathbb{Z})$, for $\mathsf{p} \in \mathsf{P}_\Gamma$ a cusp of Γ, and for $\Sigma_\mathsf{p} \in \Gamma\backslash G(\mathbb{Q})$ a scaling coset for Γ at p, we write Γ^p as a shorthand for the group $\Sigma_\mathsf{p}^{-1}\Sigma_\mathsf{p}$. In light of Lemma 6.4.1 we may reformulate the fourth condition of the main theorem of [CMS04] as follows.

Lemma 6.4.2. *Let Γ be a group of $n\|h$-type, so that $\Gamma = \Gamma_0(n\|h) + S$ for some positive integers n and h, and some subgroup S of the group of exact divisors of n/h. Then the fourth condition of the main theorem of [CMS04] is satisfied if and only if for each cusp $\mathsf{p} \in \mathsf{P}_\Gamma$ of Γ there exists a scaling coset $\Sigma_\mathsf{p} \in \Gamma\backslash G(\mathbb{Q})$ for Γ at p such that Γ^p contains $\Gamma_0(nh)$.*

We conclude this section by relating the reformulation of Lemma 6.4.2 to the normalized Rademacher sums associated to genus zero groups of $n\|h$-type.

Proposition 6.4.3. *Let Γ be a group of $n\|h$-type, so that $\Gamma = \Gamma_0(n\|h) + S$ for some positive integers n and h, and some subgroup S of the group of exact divisors of n/h, and suppose that Γ has genus zero. Let $\mathsf{p}, \mathsf{q} \in \mathsf{P}_\Gamma$ be cusps of Γ and let Σ_p and Σ_q be scaling cosets for Γ at p and q, respectively. Then the normalized Rademacher sum $T^{(-1)}_{\Gamma,\mathsf{p}|\mathsf{q}}(z)$ is $\Gamma_0(nh)$-invariant if and only if Γ^q contains $\Gamma_0(nh)$.*

Proof. Let us set $\Delta = \Gamma_0(nh)$. Observe that if we set $\mathsf{p}^\mathsf{q} = \Sigma_\mathsf{q}^{-1}\Sigma_\mathsf{p} \cdot \infty$ then p^q is a cusp of $\Gamma^\mathsf{q} = \Sigma_\mathsf{q}^{-1}\Sigma_\mathsf{q}$, and $\Sigma_\mathsf{q}^{-1}\Sigma_\mathsf{p}$ is a scaling coset for Γ^q at p^q, so the function $T^{(-1)}_{\Gamma,\mathsf{p}|\mathsf{q}}(z)$ may be identified with the normalized Rademacher sum $T^{(-1)}_{\Gamma^\mathsf{q},\mathsf{p}^\mathsf{q}}(z)$ associated to Γ^q at the cusp p^q, for $\Sigma_{\mathsf{p}^\mathsf{q}} = \Sigma_\mathsf{q}^{-1}\Sigma_\mathsf{p}$. Under the assumption that Γ has genus zero, Γ^q also has genus zero, and has width one at infinity by the defining properties (cf. Lemma 2.6.2) of Σ_q. By Theorem 6.1.3 then the normalized Rademacher sum $T^{(-1)}_{\Gamma,\mathsf{p}|\mathsf{q}}(z)$ is, up to a constant function, the expansion at p^q of the normalized hauptmodul of the group Γ^q. In particular, $T^{(-1)}_{\Gamma,\mathsf{p}|\mathsf{q}}(z)$ defines an isomorphism of Riemann surfaces $X_{\Gamma^\mathsf{q}} \to \hat{\mathbb{C}}$, and so we have $T^{(-1)}_{\Gamma,\mathsf{p}|\mathsf{q}}(\gamma \cdot z) = T^{(-1)}_{\Gamma,\mathsf{p}|\mathsf{q}}(z)$ for all $z \in \mathbb{H}$ if and only if $\gamma \in \Gamma^\mathsf{q}$. It follows then that $T^{(-1)}_{\Gamma,\mathsf{p}|\mathsf{q}}(z)$ is Δ-invariant if and only if Δ is contained in Γ^q, which is what we required to show. □

6.5 Rademacher sums and the Monster

We will now use Rademacher sums to reformulate the characterization of the groups of monstrous moonshine due to Conway–McKay–Sebbar (cf. [CMS04]).

Theorem 6.5.1. *Let Γ be a subgroup of $G(\mathbb{R})$. Then we have $\Gamma = \Gamma_g$ for some $g \in \mathbb{M}$ if and only if the following conditions are satisfied:*

- *Γ is the group defining $n\|h + S$-equivalence of solid tori for some positive integers n and h, and some subgroup $S < \mathrm{Ex}(n/h)$;*
- *the normalized Rademacher sum $T_\Gamma^{(-1)}(z)$ is Γ-invariant;*
- *there exists a system $\{\Sigma_\mathsf{p} \mid \mathsf{p} \in \mathsf{P}_\Gamma\}$ of scaling cosets for Γ such that $T_{\Gamma|\mathsf{p}}^{(-1)}(z)$ is Δ-invariant for every cusp $\mathsf{p} \in \mathsf{P}_\Gamma$ of Γ, where $\Delta = \Gamma_0(nh)$;*
- *for every cusp $\mathsf{r} \in \mathsf{P}_\Delta$ of $\Delta = \Gamma_0(nh)$ that is contained in $\Gamma\cdot\infty$ we have $T_{\Delta,\mathsf{r}}^{(-1)}(z) = T_{\Delta|\mathsf{r}}^{(-1)}(z)$ when the scaling coset Σ_r for Δ at r is taken to lie in $\Delta\backslash\Gamma$.*

Proof. According to the main theorem of [CMS04] it suffices to show that a group Γ satisfies each of the four conditions of Theorem 6.5.1 if and only if it satisfies the four conditions of the main theorem of [CMS04]. As a first step in establishing this equivalence, observe that the first condition of Theorem 6.5.1 is exactly the same as the second condition of the main theorem of [CMS04].

Suppose that Γ satisfies the four conditions of the main theorem of [CMS04]. Then, in particular, it satisfies the first condition of Theorem 6.5.1, and has genus zero. We may suppose then that $\Gamma = \Gamma_0(n\|h) + S$ for some $n, h \in \mathbb{Z}^+$, and $S < \mathrm{Ex}(n/h)$, so that Γ is commensurable with $G(\mathbb{Z})$, has width one at infinity, and contains and normalizes $\Delta = \Gamma_0(nh)$. According to Theorem 6.1.1 the normalized Rademacher sum $T_\Gamma^{(-1)}(z)$ is Γ-invariant, so the second condition of Theorem 6.5.1 is satisfied. According to Lemma 6.4.2 there exists a system $\{\Sigma_\mathsf{p} \mid \mathsf{p} \in \mathsf{P}_\Gamma\}$ of scaling cosets for Γ such that Γ^p contains Δ for each cusp $\mathsf{p} \in \mathsf{P}_\Gamma$. Applying Proposition 6.4.3 we see that the normalized Rademacher sum $T_{\Gamma|\mathsf{p}}^{(-1)}(z)$ is Δ-invariant when the scaling coset Σ_p of the above scaling coset system is chosen for the definition of $T_{\Gamma|\mathsf{p}}^{(-1)}(z)$. This confirms that the third condition of Theorem 6.5.1 is satisfied. The fourth condition of the main theorem of [CMS04] is that the quotient Γ/Δ have exponent two. Let $\mathsf{r} \in \mathsf{P}_\Delta$ be a cusp of Δ that is contained in $\Gamma\cdot\infty$, and let Σ_r be the unique right coset of Δ in Γ such that $\mathsf{r} = \Sigma_\mathsf{r}\cdot\infty$. Then the exponent two condition implies $\Sigma_\mathsf{r}^2 = \Delta$, so that the fourth condition of Theorem 6.5.1 follows from an application of Lemma 6.3.1.

Suppose now that Γ satisfies the four conditions of Theorem 6.5.1. Then Γ satisfies the second condition of [CMS04], and so we have $\Gamma = \Gamma_0(n\|h) + S$ for some $n, h \in \mathbb{Z}^+$, and $S < \mathrm{Ex}(n/h)$, and Γ is a group commensurable with $G(\mathbb{Z})$ that has width one at infinity. Applying Theorem 6.1.1 to the second condition of Theorem 6.5.1 we conclude that Γ has genus zero, and so

satisfies the first condition of [CMS04]. Set $\Delta = \Gamma_0(nh)$, so that Γ contains and normalizes Δ. Applying Proposition 6.4.3 to the third condition of Theorem 6.5.1 we see that scaling cosets $\{\Sigma_{\mathsf{p}} \mid \mathsf{p} \in \mathsf{P}_\Gamma\}$ can be chosen for Γ so that Γ^{p} contains Δ for each cusp $\mathsf{p} \in \mathsf{P}_\Gamma$. Applying Lemma 6.4.2 to this fact we conclude that Γ satisfies the fourth condition of the main theorem of [CMS04]. It remains to check that the quotient Γ/Δ has exponent two. Let $\gamma \in \Gamma$ and set $\mathsf{r} = \Delta\gamma \cdot \infty$. If we take $\Sigma_{\mathsf{r}} = \Delta\gamma$ then the fourth condition of Theorem 6.5.1 states that $T^{(-1)}_{\Delta|\mathsf{r}}(z) = T^{(-1)}_{\Delta,\mathsf{r}}(z)$. Proposition 6.3.2 now implies that $\Sigma^2_{\mathsf{r}} = \Delta$, which in turn implies $\gamma^2 \in \Delta$. This argument applies to arbitrary $\gamma \in \Gamma$ so we conclude that the quotient Γ/Δ has exponent two. This completes the proof. □

Perhaps the most technical condition of Theorem 6.5.1 is the last one. Beyond the monstrous moonshine conjectures, there are the generalized moonshine conjectures of Norton (cf. [Mas87]) which associate genus zero groups to commuting pairs of elements in the Monster. A number of the groups appearing in generalized moonshine do not satisfy the last condition of Theorem 6.5.1, but they all satisfy the second property, and a slight weakening of the first (cf. [Fer93]), and we do not know of any examples that fail to satisfy the third condition of Theorem 6.5.1. It is an interesting question then to determine if the first three conditions of Theorem 6.5.1 furnish a characterization of the groups of generalized moonshine.

7 Gravity

In this section we consider applications of the normalized Rademacher sums to chiral 3d quantum gravity. We should note that the notion of chiral 3d quantum gravity has not yet been defined, so the applications we make, and in particular the conjectures we formulate, are intended to shed light on the very problem of formulating a definition, in addition to elucidating important structural properties that a certain distinguished (and conjectural) example might satisfy. In the absence of a precise definition of 3d quantum gravity some of our discussion in this section is necessarily more speculative.

7.1 First conjecture

It is clear from our results that the Rademacher sums are particularly convenient for understanding the special characteristics of the McKay–Thompson series; in particular, their crucial genus zero property. In order to fully explain the moonshine phenomena, one has to relate these sums to the structure of the vertex operator algebra $V^\natural$. In search for the new relation, one might look again for hints from physics. Very recently, Witten, revisiting 3d quantum gravity, has formulated a number of results and observations [Wit07] including a conjecture about the existence of the 3d quantum gravities with central charges $c_L = c_R$ proportional to

24. In particular, he asserted that the simplest in his list of 3d quantum gravities should be equivalent to the 2d conformal field theory $V^\natural \otimes (V^\natural)^*$. Manschot [Man07] then suggested to consider chiral 3d quantum gravities, the simplest of which, with $c_L = 24$ and $c_R = 0$, should be equivalent exactly to $V^\natural$. Furthermore, Li, Song and Strominger [LSS08] have argued that the chiral gravity possesses stability and consistency, the necessary properties of a sound physical theory. A very recent work [MSS09] by Maloney, Song and Strominger provided further support for the existence of chiral gravities with c_L proportional to 24. These results and observations from physics taken together suggest that there exists an alternative construction of the vertex operator algebra $V^\natural$ that may be viewed as a rigorous version of the simplest chiral 3d quantum gravity with $c = 24$, in the same way as the original construction of $V^\natural$ was interpreted as a rigorous version of the chiral 2d conformal field theory with the partition function $J(z)$. In 3d quantum gravity one expects to obtain the partition function as a sum over minimum points of a 3d quantum gravity action; i.e. over all 3d hyperbolic structures on a solid torus with genus one boundary, whose conformal structure corresponds to the point z on the moduli space. Since all such structures are naturally parameterized by $\Gamma_\infty \backslash PSL_2(\mathbb{Z})$ (cf. [DMMV00]), the partition function should be a kind of Rademacher sum. In fact Manschot and Moore [MM07] argued that the subtraction of constants in (1.2.1) can be explained by a regularization of the partition function of 3d gravity. Our continued Rademacher sums of §4, leading naturally to the modified and normalized Rademacher sums, provide a good candidate for such a regularization.

To obtain the McKay–Thompson series (1.1.3) from the chiral 3d quantum gravity is a more challenging problem even at the heuristic level. However, one may look at these series from a slightly different point of view. It has been shown in [DLM00] that for any $g \in \mathbb{M}$ there exists a unique simple g-twisted $V^\natural$-module $V_g^\natural$, whose partition function is equal to $c(g)T_g(-1/z)$ where $c(g)$ is a constant depending on $g \in \mathbb{M}$. The heuristic analysis of the twisted sector $V_g^\natural$ in [Tui95] strongly supports the general assumption that $c(g) = 1$ for all $g \in \mathbb{M}$. (For more on this assumption please see the discussion of "Hypothesis A_g" in [Car09].)

We now state our first conjecture.

Conjecture 7.1.1. *There exists a family of twisted chiral 3d quantum gravities at central charge $c = 24$ associated with elements of the Monster $g \in \mathbb{M}$ whose partition functions are naturally given by sums over geometries parameterized by $\Gamma_\infty \backslash \Gamma_g$, and these partition functions coincide with the normalized Rademacher sum $T_\Gamma^{(-1)}(-1/z)$, for $\Gamma = \Gamma_g$. Moreover, the untwisted ($g = e$) chiral 3d quantum gravity has a VOA structure isomorphic to $V^\natural$, and the twisted 3d quantum gravity corresponding to $g \in \mathbb{M}$ has a structure of g-twisted $V^\natural$-module isomorphic to that of $V_g^\natural$.*

The special case $g = e$ of Conjecture 7.1.1, corresponding to the untwisted chiral 3d quantum gravity, is strongly supported by the physics literature mentioned above. We will now make a few remarks about the general twisted case. The principle of the 3d quantum gravity/2d CFT correspondence suggests that all the structures in either theory should have appropriate

counterparts in the other. The g-twisted sector $V_g^\natural$ is an intrinsic part of the extended theory of the chiral 2d CFT associated to $V^\natural$, and, according to the physical principle, should therefore have analogues in chiral 3d quantum gravity. Basic information about these twisted chiral 3d quantum gravities can be extracted from the Rademacher sums $J_g(z) = T_g(-1/z)$. We'll illustrate the general case with the level two examples, which are the cases that $\Gamma = \Gamma_0(2)$ or $\Gamma_0(2)+$. (In the notation of §2.5 the symbols $\Gamma_0(2)+$ are a shorthand for $\Gamma_0(2) + \mathrm{Ex}(2)$.) As in the case that $g = e$ one expects that each term of the Rademacher sum comes from a classical solution of the corresponding twisted chiral quantum gravity. Every Rademacher sum $J_g(z)$ contains the term $\mathrm{e}(1/z)$, which corresponds to the BTZ black hole solution (cf. [BTZ92]). The other terms arising for $\Gamma = \Gamma_0(2)$ can be characterized among all those arising for $\Gamma = G(\mathbb{Z})$ as the solutions that have the same spin structure as the BTZ black hole on the boundary. The Rademacher sum $J_g(z)$ in this case coincides with the partition function of the Ramond sector in supergravity (cf. [MW07, §7]). The group $\Gamma = \Gamma_0(2)+$ is no longer a subgroup of the modular group $G(\mathbb{Z})$. In this case we have to consider orbifold solutions in addition to smooth solutions; we consider orbifold solutions with a codimension 2 singularity along the defining circle of the solid torus which looks locally like $\mathbb{C}/\langle \mathrm{e}(1/2)\rangle$ (cf. [MW07, §2]), and we still impose the same spin structure boundary condition. Note that on any space X, choices of spin structure on X (when they exist) are in bijective correspondence with a particular family of double covers of X (cf. e.g. [LM89]). Such a double cover smooths the $\mathbb{Z}/2$-orbifold singularity. One may view the spin structure as a $\mathbb{Z}/2$-structure, and for a general level N group Γ, spin structures are replaced with $\mathbb{Z}/N$-structures. One can also allow singularities along the defining circle of the solid torus that look locally like $\mathbb{C}/\langle \mathrm{e}(1/N)\rangle$. In general, a choice of twisted gravity with $\mathbb{Z}/N$-structure group should impose an $n\|h+S$ equivalence for solid tori, for some n and h with $N = nh$ and h a divisor of 24 (cf. Theorem 6.2.3), which yields the Rademacher sum $J_g(z)$, for a corresponding $g \in \mathbb{M}$, as the saddle point approximation of the twisted quantum gravity partition function.

We expect that the conjectural twisted chiral 3d quantum gravity construction of $V_g^\natural$ will imply the Γ_g-invariance of its partition function, in analogy with the way in which 2d CFTs are found to have modular invariant partition functions (cf. [Wit87]). Then our conjecture, in combination with our results on Rademacher sums, will naturally imply the genus zero conjecture of Conway–Norton (cf. [CN79]), and most importantly, will reveal the geometric nature of moonshine. In particular, the solution to Conjecture 7.1.1 will answer the genus zero question: why the discrete groups attached to the Monster via monstrous moonshine all have genus zero.

We also expect that the conditions on Rademacher sums in our reformulation Theorem 6.5.1, of the group theoretic characterization of the groups of monstrous moonshine due to Conway–McKay–Sebbar, will also find an interpretation as consistency conditions for twisted chiral 3d quantum gravities. In particular, the completeness of the family of twisted chiral 3d quantum gravities at central charge 24 associated with elements of the Monster group $\mathbb{M}$ should have a

deep meaning in quantum gravity.

Our Conjecture 7.1.1 should have significance for the future development of analytic number theory also. We already know from 2d CFT that various constructions yield remarkable number theoretic identities. Our conjecture implies that the development of the theory of 3d quantum gravity might encompass a whole new family of number theoretic results such as the theory of Rademacher sums studied in this work. In §7.3 we'll show that the Hecke operators also admit an interpretation in terms of quantum gravity. In preparation for this we will recall in the next section some results about a certain class of generalized Kac–Moody algebras and their representations.

7.2 Monstrous Lie algebras

In his paper [Bor92], Borcherds constructed a remarkable generalized Kac–Moody algebra (*GKM algebra*) $\mathfrak{m}$, called the *Monster Lie algebra*, which plays a key rôle in his approach to monstrous moonshine. He also defined GKM superalgebras $\mathfrak{m}'_g$, for each $g \in \mathbb{M}$, by using the McKay–Thompson series $T_g(z)$ to specify the simple roots. In this section we'll consider another family of GKM algebras $\mathfrak{m}_g$, parameterized by elements $g \in \mathbb{M}$, with simple roots specified by the functions $J_g(z) = T_g(-1/z)$. Since the Fourier coefficients of the $J_g(z)$ are non-negative integers, the algebras $\mathfrak{m}_g$ are purely even. We will also study Verma modules for the GKM algebras $\mathfrak{m}_g$.

In the following section §7.3, we will explain how these algebraic structures arise from the Rademacher sums, and the conjectural twisted chiral 3d quantum gravities. This will lead us to a further extension of our first conjecture.

The algebras $\mathfrak{m}_g$ were introduced and studied by Carnahan in [Car07] and were discovered independently by the second author. We will now recall the relevant results of [Car07].

Let $L = II_{1,1}$ denote a copy of the unique even self-dual Lorentzian lattice of rank 2. We identify L with the group $\mathbb{Z} \times \mathbb{Z}$ and set the norm of the pair (m, n) to be $-2mn$. For any positive integer N we denote the sublattice $\mathbb{Z} \times N\mathbb{Z}$ by $L(N)$, and we let $L(N)^\vee$ denote the dual lattice $\frac{1}{N}\mathbb{Z} \times \mathbb{Z}$. Let V_L denote the vertex operator algebra (VOA) associated to the lattice L (cf. [FLM88]),

$$V_L = \bigoplus_{(m,n)\in L} V_L^{(m,n)}, \tag{7.2.1}$$

and let h_N denote the automorphism of V_L which acts as multiplication by $\mathrm{e}(n/N)$ on the subspace $V_L^{(m,n)}$. Then $V_{L(N)}$ is a vertex operator subalgebra of $V_{L(N)}$ fixed by h_N. Following [DL93] each coset $L + (k/N, 0)$ of L in $L(N)^\vee$ defines an h_N^k-twisted module for V_L. We denote this twisted module by $V_{L+(k/N,0)}$, and regard it as graded in the natural way by $L + (k/N, 0)$.

For $g \in \mathbb{M}$ such that the level of Γ_g is N (i.e. $N = nh$ in the notation of [CN79]) define $W_g^\natural$

to be the following $L(N)^\vee$-graded space invariant under $\langle g\rangle$.

$$W_g^\natural = \bigoplus_{k=0}^{N-1} \left(V_{g^k}^\natural \otimes V_{L+(k/N,0)} \right)^{\langle g\rangle} \tag{7.2.2}$$

The summand corresponding to $k = 0$ has a natural VOA structure, while the other summands are naturally modules for this VOA. There also exists a unique (up to scalar) intertwining operator between the product of the k-th and l-th summands and the $(k + l)$-th summand. Carnahan asserts that there is a consistent choice of these constants such that one has

Theorem 7.2.1. *The space $W_g^\natural$ naturally admits a VOA structure.*

By construction $W_g^\natural$ has rank 26 and one can define the semi-infinite cohomology of the Virasoro algebra with coefficients in $W_g^\natural$. The the first semi-infinite cohomology group acquires a Lie algebra structure according to [LZ95]. In this way we obtain a *monstrous Lie algebra* $\mathfrak{m}_g$ for each $g \in \mathbb{M}$.

$$\mathfrak{m}_g = H^{\infty/2+1}(W_g^\natural) \tag{7.2.3}$$

An alternative construction of $\mathfrak{m}_g$ may be based on the no-ghost theorem of the 26-dimensional bosonic string (cf. [Fre85], [Bor92]). For $g = e$, the Lie algebra $\mathfrak{m}_e$ is the original Monster Lie algebra of Borcherds [Bor92].

The Lie algebra $\mathfrak{m}_g$ inherits an $L(N)^\vee$-grading

$$\mathfrak{m}_g = \bigoplus_{m,n\in\mathbb{Z}} \mathfrak{m}_g^{(m/N,n)}. \tag{7.2.4}$$

The same argument as in the case $g = e$ applied to the semi-infinite cohomology or no-ghost theorem construction of $\mathfrak{m}_g$ yields natural isomorphisms

$$\mathfrak{m}_g^{(m/N,n)} \cong V_{g^{-m},n/N}^\natural(mn/N) \tag{7.2.5}$$

where $V_{g^{-m},n/N}^\natural$ is the subspace of $V_{g^{-m}}^\natural$ upon which g acts by $\mathrm{e}(n/N)$ and $V_{g^{-m},n/N}^\natural(mn/N)$ is the L_0-eigenspace of $V_{g^{-m},n/N}^\natural$ with eigenvalue $mn/N + 1$.

Remark 7.2.2. Let us write V^Λ for the vertex operator algebra associated to the Leech lattice. In his paper [Tui95] Tuite gave strong evidence that the genus zero property of monstrous moonshine is equivalent to the following duality isomorphisms,

$$V_{g^m,n/N}^\natural \cong V_{g^n,m/N}^\natural, \tag{7.2.6}$$

$$V_{g^m,n/N}^\natural \cong V_{h^n,m/N}^\Lambda, \tag{7.2.7}$$

the first holding in the case that g is Fricke, and the second in the case that g is non-Fricke. In the second case we may take g to lie in a subgroup $2^{1+24}.Co_1$ of $\mathbb{M}$ and the h in (7.2.7) denotes an element of $2^{24}.2.Co_1$, a subgroup of $\mathrm{Aut}(V^\Lambda)$, such that h and g project to the same conjugacy class of Co_1. Thus the Tuite duality in the Fricke case implies the symmetry

$$\mathfrak{m}_g^{(m/N,n)} \cong \mathfrak{m}_g^{(n/N,m)} \tag{7.2.8}$$

of the monstrous Lie algebra $\mathfrak{m}_g$, as is known to hold in the case $g = e$ studied in [Bor92]. In the non-Fricke case the Tuite duality suggests an alternative construction of the VOA $W_g^\natural$ as

$$W_g^\natural = \bigoplus_{k=0}^{N-1} \left(V_{h^k}^\Lambda \otimes V_{\tilde{L}+(0,k)} \right)^{\langle h \rangle} \tag{7.2.9}$$

where $\tilde{L} = \frac{1}{N}\mathbb{Z} \times N\mathbb{Z}$.

We now re-scale the grading in $W_g^\natural$ and $\mathfrak{m}_g$ by interchanging $(m/N, n)$ with $(m, n/N)$, and we set

$$\mathfrak{m}_g^m = \bigoplus_{n \in \frac{1}{N}\mathbb{Z}} \mathfrak{m}_g^{(m,n/N)} \tag{7.2.10}$$

for $m \in \mathbb{Z}$. Then (7.2.5) implies isomorphisms of graded spaces

$$\mathfrak{m}_g^1 \cong V_{g^{-1}}^\natural, \quad \mathfrak{m}_g^{-1} \cong V_g^\natural, \tag{7.2.11}$$

and we also have

$$\mathfrak{m}_g^0 = \mathfrak{m}_g^{(0,0)} = \mathbb{C}c \oplus \mathbb{C}d, \tag{7.2.12}$$

where c and d are the degree operators for the re-scaled grading.

In [Car07] Carnahan shows that $\mathfrak{m}_g$ is a GKM algebra, thus generalizing a result of [Bor92]. When g is Fricke, $\mathfrak{m}_g$ has a structure very similar to that of the original Borcherds Monster Lie algebra $\mathfrak{m} = \mathfrak{m}_e$. In particular, it has one real simple root and all simple roots correspond to a basis for $V_g^\natural$. This implies that $\mathfrak{m}_g$ can be reconstructed from the subalgebra

$$\mathfrak{m}_g^1 \oplus \mathfrak{m}_g^0 \oplus \mathfrak{m}_g^{-1} \tag{7.2.13}$$

which is called the *local subalgebra of* $\mathfrak{m}_g$ (cf. [FF83]).

Remark 7.2.3. When g is non-Fricke $\mathfrak{m}_g$ does not have real simple roots, and in general it cannot be reconstructed from its local subalgebra (7.2.13). However the Tuite duality suggests that if

one considers the second grading of $\mathfrak{m}_g$ then the local subalgebra consists of V_h^Λ and $V_{h^{-1}}^\Lambda$ and the h-twisted Heisenberg algebra. This local subalgebra is expected to generate $\mathfrak{m}_g$ in the case that g is non-Fricke.

In his thesis [Car07] Carnahan also obtained remarkable generalizations of the Borcherds identities for each of the monstrous Lie algebras $\mathfrak{m}_g$; viz.,

$$p(T_g(w) - J_g(z)) = \prod_{m\in\mathbb{Z}^+, n\in\mathbb{Z}} (1-p^m q^{n/N})^{c_{g^m,n/N}(mn/N)} \tag{7.2.14}$$

where we have set

$$c_{g^m,n/N}(mn/N) = \dim V^\natural_{g^m,n/N}(mn/N). \tag{7.2.15}$$

When g is Fricke the invariance of $T_g(w)$ under the Fricke involution and a re-scaling $w \mapsto Nw$ yields

$$p(J_g(Nw) - J_g(Nz)) = \prod_{m\in\mathbb{Z}^+, n\in\mathbb{Z}} (1-p^m q^{n})^{c_{g^m,n/N}(mn/N)}. \tag{7.2.16}$$

Dividing both sides of (7.2.16) by $(1-pq^{-1})$ we obtain expressions which are invariant under the transposition of p with q. This entails the identity

$$c_{g^m,n/N}(mn/N) = c_{g^m,m/N}(mn/N) \tag{7.2.17}$$

which is in agreement with the Tuite duality of (7.2.6). The identity (7.2.14) implies an unexpected formula for the graded dimension of the Verma module with trivial character $\mathcal{V}_g$ associated to g.

$$\mathcal{V}_g = \mathcal{U}(\mathfrak{m}_g^-) \tag{7.2.18}$$

In fact the standard product formula

$$\operatorname{gdim} \mathcal{V}_g = \prod_{m\in\mathbb{Z}^+, n\in\mathbb{Z}} (1-p^m q^{n/N})^{-c_{g^m,n/N}(mn/N)} \tag{7.2.19}$$

for the graded dimension of $\mathcal{V}_g$ follows from (7.2.5) and (7.2.15). The identity (7.2.14) yields the alternative expression

$$\operatorname{gdim} \mathcal{V}_g = \frac{1}{p(T_g(w) - J_g(z))}. \tag{7.2.20}$$

This expression allows us to view the bi-graded dimension as a meromorphic function on $\mathbb{H}\times\mathbb{H}$.

Thanks to the properties of principal moduli for curves of genus zero, the expression (7.2.20) is singular at the point (w, z) precisely when $\Gamma_g \cdot w = \Gamma_g \cdot (-1/z)$. In §7.3 we will give an interpretation of this fact in the setting of three dimensional quantum gravity.

With the goal of making such an interpretation in mind we propose to rewrite the first expression (7.2.19) for the bi-graded dimension of the Verma module $\mathcal{V}_g$ as

$$\operatorname{gdim} \mathcal{V}_g = \exp\left(\sum_{m\in\mathbb{Z}^+} \hat{T}_g(m) J_g(z) p^m\right) \tag{7.2.21}$$

using generalized Hecke operators $\hat{T}_g(m)$ such as those discussed in §5.5; we have verified there the existence of such operators for $g = e$ and we conjecture that suitable $\hat{T}_g(m)$ exist for every $g \in \mathbb{M}$. Comparing then with the generating functions

$$Z_g(p, q) = \sum_{m\in\mathbb{Z}^+} m\hat{T}_g(m) J_g(z) p^m \tag{7.2.22}$$

of the higher order Rademacher sums $J_g^{(-m)}(z)$ we obtain the following expression.

$$Z_g(p, q) = p\partial_p \log(\operatorname{gdim} \mathcal{V}_g) \tag{7.2.23}$$

An identity of this form is well-known in the theory of symmetric functions (cf. [Mac95]), where it serves to relate the generating function of the power symmetric functions

$$p_n = \sum_{1\le i} x_i^n \tag{7.2.24}$$

with that of the complete symmetric functions

$$h_n = \sum_{1\le i_1\le\cdots\le i_n} \prod_{k=1}^{n} x_{i_k}. \tag{7.2.25}$$

Thus the graded dimensions of the Verma modules $\mathcal{V}_g$ may be viewed as *complete Rademacher sums*, where the rôle of the variables x_i, for $i \in \mathbb{Z}^+$, in (7.2.24) and (7.2.25) is taken up by the exponential expressions $\mathrm{e}([\![\gamma \cdot z']\!])$ for $z' = -1/z$ and $[\![\gamma \in [\![\Gamma_g$.

Since all the coefficients appearing in the bi-graded dimension $Z_g(p, q)$ are non-negative integers, one may expect to naturally find a bi-graded vector space, and possibly even an $\mathfrak{m}_g$-module, with bi-graded dimension given by $Z_g(p, q)$. The next result suggests that one can expect to find such a space within $\mathcal{V}_g$.

Given two elements $F = \sum F_{m,n} p^m q^n$ and $G = \sum G_{m,n} p^m q^n$ in $\mathbb{Z}((q))[[p]]$, let us write $F \le G$ in the case that $F_{m,n} \le G_{m,n}$ for all $m, n \in \mathbb{Z}$.

Proposition 7.2.4. *Let $g \in \mathbb{M}$. If g is Fricke then we have $Z_g(p,q) \leq \operatorname{gdim} \mathcal{V}_g$.*

Proof. Dividing both sides of (7.2.16) by $(1-pq^{-1})$ we obtain

$$1 - \sum_{m,n\in\mathbb{Z}^+} c_{g,\frac{m+n-1}{N}}(\frac{m+n-1}{N})p^m q^n = \prod_{m,n\in\mathbb{Z}^+}(1-p^m q^n)^{c_g{}^{m,n/N}(mn/N)}. \tag{7.2.26}$$

We denote the sum in the left hand side by Σ^+ and the product in the right hand side by Π^+. Then we have

$$Z_g(p,q) = -p\partial_p \log(1-pq^{-1})\Pi^+ = \frac{pq^{-1}}{1-pq^{-1}} + \frac{p\partial_p \Sigma^+}{1-\Sigma^+}. \tag{7.2.27}$$

On the other hand

$$\operatorname{gdim} \mathcal{V}_g = \frac{1}{1-pq^{-1}}\frac{1}{1-\Sigma^+} = \frac{1}{1-pq^{-1}} + \frac{\Sigma^+}{(1-pq^{-1})(1-\Sigma^+)}. \tag{7.2.28}$$

Thus it is sufficient to show that

$$p\partial_p(\Sigma^+) \leq \frac{1}{1-pq^{-1}}\Sigma^+, \tag{7.2.29}$$

but this follows from the observation that both sides of (7.2.29) have nonnegative coefficients and the left hand side may be obtained from the right hand side by deleting the terms which are singular or constant with respect to q. □

To formulate a conjecture on the nature of the subspaces of $\mathcal{V}_g$ with bi-graded dimensions $Z_g(p,q)$, as well as the origin of the higher order Rademacher sums, we turn again, in the next section, to the structures of twisted chiral 3d quantum gravity.

7.3 Second conjecture

We have shown in §5.5 that the higher order Rademacher sums recover the action of the Hecke operators on the first order Rademacher sums. In view of the conjectural relation between first order Rademacher sums and chiral 3d quantum gravities at central charge $c = 24$, we may consider an analogous interpretation for the higher order Rademacher sums given by $m(\hat{T}_g(m)J_g)(z)$ for $g \in \mathbb{M}$ and $m \in \mathbb{Z}^+$. Since all the Fourier coefficients of these expressions are non-negative integers, one might guess that there exist families of twisted chiral 3d quantum gravities for all central charges $c = 24m$, for $m \in \mathbb{Z}^+$. In particular, the untwisted ($g = e$) chiral 3d quantum gravities possess the structure of certain extremal vertex algebras with the Monster symmetry.

The aforementioned class of extremal vertex algebras, with the addition of a Virasoro element, has been originally conjectured by Witten [Wit07]. However, it was shown in [Gai08] and

[Hoe08] that the addition of a Virasoro element precludes a non-trivial action of the Monster group. The higher order Rademacher sums also point to the partition functions without elements of spin 2. We may modify Witten's conjecture by asking for the existence of extremal vertex algebras with partition functions given by the function $J^{(-m)}(z)$ for $m > 0$, related to the higher order Rademacher sums via $J^{(-m)}(z) = Q_\Gamma^{(-m)}(z) - \mathrm{c}_\Gamma(-m, 0)$, with $\Gamma = PSL_2(\mathbb{Z})$. The validity of this conjecture remains open.

Note that the Virasoro algebra, even when not represented by any actual state, may still act on a vertex algebra with partition function $J^{(-m)}(z)$. Regardless of whether or not spaces $V^{(-m)}$ with partition function $J^{(-m)}(z)$ admit natural vertex algebra structures, they certainly naturally inherit actions of the Monster group, since the actions of Hecke operators $m\hat{T}(m)$ on the Fourier coefficients of $J(z)$ may be interpreted as actions on representations of $\mathbb{M}$ (cf. [JLW95]).

It is natural to enquire as to the meaning of the spaces $V_g^{(-m)}$, for $m \in \mathbb{Z}^+$ and $g \in \mathbb{M}$, with graded dimension $J_g^{(-m)}(z)$. In light of our First Conjecture, it is unlikely that the spaces $V^{(-m)}$ can represent states of a three dimensional quantum gravity with $c = 24m$, since a Virasoro element is generally not present. Thus we have to conclude that the three dimensional quantum gravity $V^{(-1)} = V^\natural$ is the only viable candidate. Then one might view the spaces $V^{(-m)}$ for $m > 1$ as some higher overtones of the basic $m = 1$ theory.

In fact, we have shown in §5.5 that the action of the classical Hecke operator $\hat{T}(m)$ on the Rademacher sum (1.2.1) yields a sum over $\Gamma_\infty \backslash M(m)$, where $M(m)$ denotes the (image in $G(\mathbb{Q})$ of the) set of 2×2 matrices with integral entries and determinant m. This sum may be viewed as the *m-instanton correction* of the partition function in three dimensional quantum gravity. By *n-instanton* in this context one understands an elliptic curve which admits a holomorphic map of degree n into a given elliptic curve $E_{\mathbb{I}z}$ (cf. [DMVV97]). Thus the sum over $\Gamma_\infty \backslash M(n)$ becomes a sum over all three dimensional hyperbolic structures on a solid torus with genus one boundary whose conformal structure corresponds to an n-instanton on $E_{\mathbb{I}z}$. The same phenomena is expected for all the twisted 3d quantum gravities corresponding to elements $g \in \mathbb{M}$.

Collecting all the m-instanton contributions in one generating function we then obtain a partition function

$$Z_g(p, q) = \sum_{m \in \mathbb{Z}^+} m(\hat{T}_g(m) J_g)(z) p^m \tag{7.3.1}$$

depending on two variables, for each $g \in \mathbb{M}$. This partition function $Z_g(p, q)$ may be viewed as a part of the full partition function

$$\tilde{Z}_g(p, q) = \exp\left(\sum_{m \in \mathbb{Z}^+} (\hat{T}_g(m) J_g)(z) p^m \right) \tag{7.3.2}$$

of the stringy second quantization of $V_g^\natural$, introduced in [DMVV97] in the untwisted case $g = e$. (See also [Tui08].) We will now generalize their construction to an arbitrary twisted module $V_g^\natural$.

Let V be an arbitrary vertex operator algebra; in our case $V = V^\natural$. For n a positive integer, let $V^{\otimes n}$ denote the tensor product of n copies of V, and let Z_n be the group of cyclic permutations of the factors generated by an n-cycle $\sigma_n = (1, 2, \cdots, n)$. Let g be an automorphism of V of order N, then $g \times \sigma_n$ is an automorphism of $V^{\otimes n}$ of order nN/d where $d = (n, N)$. To the pair $(V^{\otimes n}, g \times \sigma_n)$ is canonically associated a twisted module (cf. [BDM02]), which we denote $V_{(g,n)}$. The group $Z_{g,n} = \langle g \times \sigma_n \rangle$ acts naturally on $V_{(g,n)}$. We write $V_{(g,n)}^{Z_{g,n}}$ for the $Z_{g,n}$-invariant subspace. The *stringy second quantization of the twisted module* V_g, to be denoted $\mathbb{S}\, V_g$, is, by definition, the space

$$\mathbb{S}\, V_g = \bigoplus_{\lambda \in \mathcal{P}} \bigotimes_{r>0} S^{m_r} V_{(g,r)}^{Z_{g,r}}, \tag{7.3.3}$$

where the sum is taken over all partitions λ with m_r parts of length $r > 0$. The space $\mathbb{S}\, V_g$ is doubly graded: by the degrees of products of twisted sectors and by the value of $|\lambda|$. For $g = e$ our definition of the stringy second quantization coincides with that of [DMVV97].

Next we establish a relation between the second quantization $\mathbb{S}\, V_g^\natural$ and the Verma module $\mathcal{V}_g$. First note the canonical isomorphism

$$\mathcal{U}(\mathfrak{m}_g^-) \cong S(\mathfrak{m}_g^-). \tag{7.3.4}$$

Then we obtain the following result.

Theorem 7.3.1. *There is a canonical isomorphism of bi-graded vector spaces*

$$\mathbb{S}\, V_g^\natural \cong S(\mathfrak{m}_g^-). \tag{7.3.5}$$

Proof. The isomorphism (7.3.5) follows from the existence of isomorphisms

$$(V_{(g,n)}^\natural)^{Z_{g,n}} \cong \mathfrak{m}_g^{-n} \tag{7.3.6}$$

of bi-graded vector spaces for each $n \in \mathbb{Z}^+$. To establish the existence of these we will use the isomorphism

$$V_{(g,n)}^\natural \cong V_{g^n}^\natural \tag{7.3.7}$$

of [BDM02].

We wish to show that the subspace of $V_{g^n}^\natural$ fixed by the operator

$$g \times \mathrm{e}((L_0 - 1)/n) \tag{7.3.8}$$

is naturally isomorphic to $\mathfrak{m}_g^n$. Consider the invariant subspace of $V_{g^n}^\natural$ with respect to the N-th power of the operator (7.3.8), which may be expressed as $\mathrm{e}((L_0 - 1)N/n)$. Its action on the graded subspace $V_{g^n}^\natural(kd/N)$ for $k \in \mathbb{Z}$ is multiplication by $\mathrm{e}(kd/n)$, and this scalar is 1 if and only if $k = (n/d)m$ for some $m \in \mathbb{Z}$. Thus we are only concerned with the graded subspaces $V_{g^n}^\natural(nm/N)$ where $m \in \mathbb{Z}$. To find the invariant subspaces with respect to the operator (7.3.8) we consider its action on the subspaces of the form $V_{g^n,m'/N}^\natural(mn/N)$. The action is scalar multiplication by $\mathrm{e}(-m'/N)\,\mathrm{e}(m/N)$ and is therefore trivial if and only if $m \equiv m' \pmod N$. Thus the full subspace of $V_{g^n}^\natural$ invariant under (7.3.8) is given by

$$\bigoplus_{m\in\mathbb{Z}} V_{g^n,m/N}^\natural(mn/N) \cong \mathfrak{m}_g^{-n}. \tag{7.3.9}$$

To complete the proof we note that the group generated by the action of the operator (7.3.8) has order nN/d, and under the isomorphism (7.3.7) its action recovers that of $Z_{g,n}$. □

The isomorphism of Theorem 7.3.1 suggests that the stringy second quantization $\mathbb{S}\,V_g^\natural$ admits an action by the Lie algebra $\mathfrak{m}_g$ and thus provides a Fock space realization of the Verma module $\mathcal{V}_g$. When $\mathfrak{m}_g$ is generated by its local subalgebra (7.2.13), as happens in the Fricke case, it is sufficient to describe the action of brackets $[\mathfrak{m}_g^{\pm 1}, \mathfrak{m}_g^{-n}] \subset \mathfrak{m}_g^{-n\pm 1}$ via the action of $V_{g^{\pm 1}}^\natural$ on the twisted sector for $((V^\natural)^{\otimes n}, g\times\sigma_n)$. Note also that the removal of the subspaces corresponding to the terms which are singular or constant with respect to q, as in the proof of Proposition 7.2.4, will yield an embedding of the instanton subspace $\mathcal{I}_g$ inside the stringy second quantization of $V_g^\natural$.

Our interpretation of the higher order Rademacher sums and the complete Rademacher sums, and their relation to the second quantization of the moonshine vertex operator algebra and its twisted modules, as well as the monstrous Lie algebras and their Verma modules, suggests the following extension of our first conjecture, Conjecture 7.1.1.

Conjecture 7.3.2. *Assuming the existence of a family of g-twisted chiral 3d quantum gravities for $g \in \mathbb{M}$, having the properties stated in Conjecture 7.1.1, there also exists a natural geometric interpretation of the family of twisted monstrous Lie algebras $\mathfrak{m}_g$, the instanton spaces $\mathcal{I}_g$, and the denominator formulas for all $g \in \mathbb{M}$ via the second quantization of the corresponding g-twisted chiral three 3d quantum gravities.*

First of all, the partition function of the second quantized g-twisted chiral 3d quantum gravity should depend on two modular parameters w and z, and should respect the symmetry

that interchanges w with $-1/z$. Second, there should be a quantum gravity theoretic explanation for why these partition functions are singular precisely when w and $-1/z$ belong to the same orbit of Γ_g. This will explain the remarkable denominator formulas (7.2.14) of Carnahan, and ultimately, the fundamental rôle of the principal moduli in the moonshine Conjectures.

Where can we find an appropriate setting for all the structures that appear in our Second Conjecture? The first answer that might come to mind is the 26 dimensional bosonic string theory. It was known for a long time that the physical space has a Lie algebra structure (cf. [Fre85], [GO85]), which can also be recovered from the semi-infinite cohomology (cf. [FGZ86], [LZ95]). However, in this way one can only get a fake Monster Lie algebra (cf. [Bor92]) and there is no (straight forward) geometric way to turn it into the real Monster Lie algebra that appears in the three dimensional quantum gravity approach. Also, the 26 dimensions of the bosonic string have nothing to do with the 3 dimensions of the quantum gravity. It is still possible that some constructions of string theory can be applied in the 3d quantum gravity setting. In particular, one can expect to identify the negative part $\mathfrak{m}^-$ of the Monster Lie algebra in the space of the second quantization of $V^\natural$ with a certain version of BPS states (cf. [HM96]).

Clearly, our results on Rademacher sums, their relation to the moonshine module, monstrous Lie algebras, and 3d quantum gravity, admit a super-counterpart [DF09]. In this case the promise of relationships with the structures of the 10 dimensional superstring is even more tempting, since the latter has played such a prominent rôle in physics over the past 25 years. However, in spite of the of the remarkable mathematical similarity, the 10 of the ten dimensional superstring and the 3 of the 3d quantum supergravity emerging from the super-counterpart of the Rademacher sums have different geometric meaning and are not related by a compactification of any kind.

The most fascinating fact about the 3d quantum supergravity is that while it is similar, but not directly related, to the initial 3d quantum gravity, the corresponding second quantized theories are directly related. The higher order Rademacher sums point to a larger second quantized space in the super-case; a space which contains the second quantized space we considered above.

Then what can these second quantized 3d quantum gravities mean in physics? Naïvely, any quantization can be interpreted as a categorification, which lifts a given theory one dimension up. One can then wonder if the second quantization in our case might point to an extreme sector of certain four dimensional quantum gravities, such as the extreme Kerr black hole, which was recently found in [GHSS08] also to be dual to a chiral 2d CFT. In this case the Monster (or moonshine) might be the answer to the perpetual question of what is behind the letter M in the theory that has not yet revealed its true name.

Acknowledgement

We are very grateful to the referees for many helpful comments and suggestions.

References

[AS64] Milton Abramowitz and Irene A. Stegun. *Handbook of mathematical functions with formulas, graphs, and mathematical tables*, volume 55 of *National Bureau of Standards Applied Mathematics Series*. For sale by the Superintendent of Documents, U.S. Government Printing Office, Washington, D.C., 1964.

[BDM02] Katrina Barron, Chongying Dong, and Geoffrey Mason. Twisted sectors for tensor product vertex operator algebras associated to permutation groups. *Comm. Math. Phys.*, 227(2):349–384, 2002.

[Bor92] Richard E. Borcherds. Monstrous moonshine and monstrous Lie superalgebras. *Invent. Math.*, 109(2):405–444, 1992.

[BTZ92] Máximo Bañados, Claudio Teitelboim, and Jorge Zanelli. Black hole in three-dimensional spacetime. *Phys. Rev. Lett.*, 69(13):1849–1851, 1992.

[Car07] Scott Carnahan. *Monstrous Lie algebras and generalized moonshine*. PhD thesis, University of California, Berkeley, 2007.

[Car09] Scott Carnahan. Generalized Moonshine II: Borcherds products. arXiv:0908.4223.

[CMS04] John Conway, John McKay, and Abdellah Sebbar. On the discrete groups of Moonshine. *Proc. Amer. Math. Soc.*, 132:2233–2240, 2004.

[CN79] J. H. Conway and S. P. Norton. Monstrous moonshine. *Bull. London Math. Soc.*, 11(3):308–339, 1979.

[DF09] John F. Duncan and Igor B. Frenkel. Rademacher sums and super-moonshine. In preparation.

[DL93] Chongying Dong and James Lepowsky. *Generalized vertex algebras and relative vertex operators*, volume 112 of *Progress in Mathematics*. Birkhäuser Boston Inc., Boston, MA, 1993.

[DLM00] Chongying Dong, Haisheng Li, and Geoffrey Mason. Modular-invariance of trace functions in orbifold theory and generalized Moonshine. *Communications in Mathematical Physics*, 214:1–56, 2000.

[DMMV00] Robbert Dijkgraaf, Juan Maldacena, Gregory Moore, and Erik Verlinde. A black hole Farey tail. arXiv:hep-th/0005003v3, April 2000.

[DMVV97] Robbert Dijkgraaf, Gregory Moore, Erik Verlinde, and Herman Verlinde. Elliptic genera of symmetric products and second quantized strings. *Comm. Math. Phys.*, 185(1):197–209, 1997.

[Fer93] Charles R. Ferenbaugh. The genus-zero problem for $n|h$-type groups. *Duke Math. J.*, 72(1):31–63, 1993.

[FF83] Alex J. Feingold and Igor B. Frenkel. A hyperbolic Kac-Moody algebra and the theory of Siegel modular forms of genus 2. *Math. Ann.*, 263(1):87–144, 1983.

[FGZ86] I. B. Frenkel, H. Garland, and G. J. Zuckerman. Semi-infinite cohomology and string theory. *Proc. Nat. Acad. Sci. U.S.A.*, 83(22):8442–8446, 1986.

[FLM88] Igor Frenkel, James Lepowsky, and Arne Meurman. *Vertex operator algebras and the Monster*, volume 134 of *Pure and Applied Mathematics*. Academic Press Inc., Boston, MA, 1988.

[Fre85] I. B. Frenkel. Representations of Kac-Moody algebras and dual resonance models. In *Applications of group theory in physics and mathematical physics (Chicago, 1982)*, volume 21 of *Lectures in Appl. Math.*, pages 325–353. Amer. Math. Soc., Providence, RI, 1985.

[Gai08] Davide Gaiotto. Monster symmetry and extremal CFTs. arXiv:0801.0988v1, January 2008.

[GHSS08] Monica Guica, Thomas Hartman, Wei Song, and Andrew Strominger. The Kerr/CFT correspondence. arXiv:0809.4266v1, September 2008.

[GO85] P. Goddard and D. Olive. Algebras, lattices and strings. In *Vertex operators in mathematics and physics (Berkeley, Calif., 1983)*, volume 3 of *Math. Sci. Res. Inst. Publ.*, pages 51–96. Springer, New York, 1985.

[GS83] D. Goldfeld and P. Sarnak. Sums of Kloosterman sums. *Invent. Math.*, 71(2):243–250, 1983.

[Hej83] Dennis A. Hejhal. *The Selberg trace formula for* PSL(2, $\mathbf{R}$). *Vol. 2*, volume 1001 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 1983.

[HM96] Jeffrey A. Harvey and Gregory Moore. Algebras, BPS states, and strings. *Nuclear Phys. B*, 463(2-3):315–368, 1996.

[Hoe08] G. Hoehn. Self-dual vertex operator superalgebras of large minimal weight. arXiv:0801.1822v1, January 2008.

[HR64] G. H. Hardy and M. Riesz. *The general theory of Dirichlet's series*. Cambridge Tracts in Mathematics and Mathematical Physics, No. 18. Stechert-Hafner, Inc., New York, 1964.

[Iwa97] Henryk Iwaniec. *Topics in classical automorphic forms*, volume 17 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 1997.

[Iwa02] Henryk Iwaniec. *Spectral methods of automorphic forms*, volume 53 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, second edition, 2002.

[JLW95] E. Jurisich, J. Lepowsky, and R. L. Wilson. Realizations of the Monster Lie algebra. *Selecta Math. (N.S.)*, 1(1):129–161, 1995.

[Kno61a] Marvin Isadore Knopp. Construction of a class of modular functions and forms. *Pacific J. Math.*, 11:275–293, 1961.

[Kno61b] Marvin Isadore Knopp. Construction of a class of modular functions and forms. II. *Pacific J. Math.*, 11:661–678, 1961.

[Kno62a] Marvin Isadore Knopp. Construction of automorphic forms on H-groups and supplementary Fourier series. *Trans. Amer. Math. Soc.*, 103:168–188, 1962.

[Kno62b] Marvin Isadore Knopp. On abelian integrals of the second kind and modular functions. *Amer. J. Math.*, 84:615–628, 1962.

[Kno86] M. I. Knopp. On the Fourier coefficients of small positive powers of $\theta(\tau)$. *Invent. Math.*, 85(1):165–183, 1986.

[Kno89] Marvin I. Knopp. On the Fourier coefficients of cusp forms having small positive weight. In *Theta functions—Bowdoin 1987, Part 2 (Brunswick, ME, 1987)*, volume 49 of *Proc. Sympos. Pure Math.*, pages 111–127. Amer. Math. Soc., Providence, RI, 1989.

[Kno90] Marvin I. Knopp. Rademacher on $J(\tau)$, Poincaré series of nonpositive weights and the Eichler cohomology. *Notices Amer. Math. Soc.*, 37(4):385–393, 1990.

[KR01] Marvin Knopp and Sinai Robins. Easy proofs of Riemann's functional equation for $\zeta(s)$ and of Lipschitz summation. *Proc. Amer. Math. Soc.*, 129(7):1915–1922 (electronic), 2001.

[Kub73] Tomio Kubota. *Elementary theory of Eisenstein series*. Kodansha Ltd., Tokyo, 1973.

[LM89] H. Blaine Lawson, Jr. and Marie-Louise Michelsohn. *Spin geometry*, volume 38 of *Princeton Mathematical Series*. Princeton University Press, Princeton, NJ, 1989.

[LSS08] Wei Li, Wei Song, and Andrew Strominger. Chiral gravity in three dimensions. *J. High Energy Phys.*, (4):082, 15, 2008.

[LZ95] Bong H. Lian and Gregg J. Zuckerman. Moonshine Cohomology. arXiv:q-alg/9501015v1, January 1995.

[Mac95] I. G. MacDonald. *Symmetric functions and Hall polynomials*. Oxford Mathematical Monographs. The Clarendon Press Oxford University Press, New York, second edition, 1995. With contributions by A. Zelevinsky, Oxford Science Publications.

[Man07] Jan Manschot. AdS_3 partition functions reconstructed. *J. High Energy Phys.*, (10):103, 7, 2007.

[Mas87] Geoffrey Mason. Finite groups and modular functions. In *The Arcata Conference on Representations of Finite Groups (Arcata, Calif., 1986)*, volume 47 of *Proc. Sympos. Pure Math.*, pages 181–210. Amer. Math. Soc., Providence, RI, 1987. With an appendix by S. P. Norton.

[McM92] Curt McMullen. Riemann surfaces and the geometrization of 3-manifolds. *Bull. Amer. Math. Soc. (N.S.)*, 27(2):207–216, 1992.

[MM07] Jan Manschot and Gregory Moore. A modern Farey tail. arXiv:0712.0573v1, December 2007.

[MSS09] Alexander Maloney, Wei Song, and Andrew Strominger. Chiral gravity, log gravity and extremal CFT. arXiv:0903.4573v1, March 2009.

[MW07] A. Maloney and E. Witten. Quantum gravity partition functions in three dimensions. arXiv:0712.0155v1, December 2007.

[Nie73] Douglas Niebur. A class of nonanalytic automorphic functions. *Nagoya Math. J.*, 52:133–145, 1973.

[Nie74] Douglas Niebur. Construction of automorphic forms and integrals. *Trans. Amer. Math. Soc.*, 191:373–385, 1974.

[Nor84] S. P. Norton. More on moonshine. In *Computational group theory (Durham, 1982)*, pages 185–193. Academic Press, London, 1984.

[Pri00] Wladimir de Azevedo Pribitkin. A generalization of the Goldfeld-Sarnak estimate on Selberg's Kloosterman zeta-function. *Forum Math.*, 12(4):449–459, 2000.

[Rad38] Hans Rademacher. The Fourier Coefficients of the Modular Invariant $J(\tau)$. *Amer. J. Math.*, 60(2):501–512, 1938.

[Rad39] Hans Rademacher. The Fourier Series and the Functional Equation of the Absolute Modular Invariant $J(\tau)$. *Amer. J. Math.*, 61(1):237–248, 1939.

[Sel65] Atle Selberg. On the estimation of Fourier coefficients of modular forms. In *Proc. Sympos. Pure Math., Vol. VIII*, pages 1–15. Amer. Math. Soc., Providence, R.I., 1965.

[Sul81] Dennis Sullivan. On the ergodic theory at infinity of an arbitrary discrete group of hyperbolic motions. In *Riemann surfaces and related topics: Proceedings of the*

1978 Stony Brook Conference (State Univ. New York, Stony Brook, N.Y., 1978), volume 97 of *Ann. of Math. Stud.*, pages 465–496. Princeton Univ. Press, Princeton, N.J., 1981.

[Tho79a] J. G. Thompson. Finite groups and modular functions. *Bull. London Math. Soc.*, 11(3):347–351, 1979.

[Tho79b] J. G. Thompson. Some numerology between the Fischer-Griess Monster and the elliptic modular function. *Bull. London Math. Soc.*, 11(3):352–353, 1979.

[Tui95] Michael P. Tuite. On the relationship between Monstrous Moonshine and the uniqueness of the Moonshine module. *Comm. Math. Phys.*, 166(3):495–532, 1995.

[Tui08] Michael Tuite. Monstrous and Generalized Moonshine and permutation orbifolds. arXiv:0811.4525v1, November 2008.

[Wit87] Edward Witten. Physics and geometry. In *Proceedings of the International Congress of Mathematicians, Vol. 1, 2 (Berkeley, Calif., 1986)*, pages 267–303, Providence, RI, 1987. Amer. Math. Soc.

[Wit07] Edward Witten. Three-dimensional gravity revisited. arXiv:0706.3359v1, June 2007.

What is the monster.

Richard E. Borcherds, *

Mathematics department, Evans Hall #3840, University of California at Berkeley, CA 94720-3840 U. S. A.

e-mail: reb@math.berkeley.edu

www home page www.math.berkeley.edu/~reb

arXiv:math/0209328v1 [math.GR] 24 Sep 2002

When I was a graduate student, my supervisor John Conway would bring his 1 year old son into the department, who was soon known as the baby monster. A more serious answer to the question is that the monster is the largest of the (known**) sporadic simple groups. Its name comes from its size: the number of elements is

$$\begin{aligned}&8080,17424,79451,28758,86459,90496,17107,57005,75436,80000,00000\\ =&2^{46}.3^{20}.5^{9}.7^{6}.11^{2}.13^{3}.17.19.23.29.31.41.47.59.71,\end{aligned}$$

about equal to the number of elementary particles in the planet Jupiter.

The monster was originally predicted to exist by Fischer and by Griess in the early 1970's. Griess constructed it a few years later in an extraordinary tour de force as the group of linear transformations on a vector space of dimension 196883 that preserved a certain commutative but non-associative bilinear product, now called the Griess product.

Our knowledge of the structure and representations of the monster is now pretty good. The 194 irreducible complex representations were worked out by Fischer, Livingstone and Thorne (before the monster was even shown to exist) They take up 8 large pages in the atlas [A] of finite groups, which is the best single source of information about the monster (and other finite simple groups). The subgroup structure is mostly known; in particular there is an almost complete list of the maximal subgroups, and the main gaps in our knowledge concern embeddings of very small simple groups in the monster. If anyone wishes to multiply elements of the monster explicitly, R. A. Wilson can supply two matrices that generate the monster. But there is a catch: each matrix takes up about 5 gigabytes of storage, and to quote from Wilson's atlas page: "standard generators have now been made as 196882×196882 matrices over GF(2)... They have been multiplied together, using most of the computing resources of Lehrstuhl D für Mathematik, RWTH Aachen for about 45 hours...". (The difficulty of multiplying two elements of the monster is caused not so much by its huge size as by the lack of "small" representations; for example, the symmetric group S_{50} is quite a lot bigger than the monster, but it only takes a few minutes to multiply two elements by hand.) Finally the modular representations of the monster for large primes were worked out by Hiss and Lux; the ones for small primes still seem to be out of reach at the moment.

* Supported by NSF grant DMS-9970611.

** The announcement of the classification of the finite simple groups about 20 years ago was a little over-enthusiastic, but a recent 1300 page preprint by Aschbacher and Smith should finally complete it.

In the late 1970's John McKay decided to switch from finite group theory to Galois theory. One function that turns up in Galois theory is the elliptic modular function

$$j(\tau) = q^{-1} + 744 + 196884q + 21493760q^2 + \cdots = \sum c(n)q^n$$

($q = e^{2\pi i\tau}$), which is essentially the simplest non-constant function invariant under the action $\tau \mapsto (a\tau + b)/(c\tau + d)$ of $SL_2(\mathbf{Z})$ on the upper half plane $\{\tau|\Im(\tau) > 0\}$. He noticed that the coefficient 196884 of q^1 was almost equal to the degree 196883 of the smallest complex representation of the monster (up to a small experimental error). The term "moonshine" roughly means weird relations between sporadic groups and modular functions (and anything else) similar to this. It was clear to many people that this was just a meaningless coincidence; after all, if you have enough large integers from various areas of mathematics then a few are going to be close just by chance, and John McKay was told that his observation was about as useful as looking at tea-leaves. John Thompson took McKay's observation further, and pointed out that the next few coefficients of the elliptic modular function were also simple linear combinations of dimensions of irreducible representations of the monster; for example, $21493760 = 21296876 + 196883 + 1$. He suggested that there should be a natural infinite dimensional graded representation $V = \sum_{n\in\mathbf{Z}} V_n$ of the monster such that the dimension of V_n is the coefficient $c(n)$ of q^n in $j(\tau)$, at least for $n \neq 0$. (The constant term of $j(\tau)$ is arbitrary as adding a constant to j still produces a function invariant under $SL_2(\mathbf{Z})$, and is set equal to 744 mainly for historical reasons.) Conway and Norton [C-N] followed up Thompson's suggestion of looking at the McKay-Thompson series $T_g(\tau) = \sum_n Trace(g|V_n)q^n$ whose coefficients are given by the traces of elements g of the monster on the representations V_n, and found by calculating the first few terms that they all seemed to be Hauptmoduls of genus 0. (A Hauptmodul is a function similar to j, but invariant under some group other than $SL_2(\mathbf{Z})$). Atkin, Fong, and Smith showed by computer calculation that there was indeed an infinite dimensional graded representation of the monster whose McKay-Thompson series were the Hauptmoduls found by Conway and Norton, and soon afterwards Frenkel, Lepowsky, and Meurman explicitly constructed this representation using vertex operators.

If a group acts on a vector space it is natural to ask if it preserves any algebraic structure, such as a bilinear form or product. The monster module constructed by FLM has a vertex algebra structure invariant under the action of the monster. Unfortunately there is no easy way to explain what a vertex algebra is; see [K] for the best introduction to them. Vertex algebras are a generalization of commutative rings with derivations (at least in characteristic 0). Roughly speaking they can be thought of as commutative rings with derivation where the ring multiplication is not quite defined everywhere; this is analogous to rational maps in algebraic geometry, which are also not quite defined everywhere. A more concrete but less intuitive definition of a vertex algebra is that it consists of a space with a countable number of bilinear products satisfying certain rather complicated identities. In the case of the monster vertex algebra $V = \oplus V_n$, this gives bilinear maps from $V_i \times V_j$ to V_k for all integers i, j, k, and the special case of the map from $V_2 \times V_2$ to V_2 is (essentially) the Griess product. So the Griess algebra is a sort of section of the monster vertex algebra.

Following an idea of Frenkel, the monster vertex algebra can be used to construct the monster Lie algebra by using the Goddard-Thorn no-ghost theorem from string theory.

This is a $\mathbf{Z}^2$-graded Lie algebra, whose piece of degree $(m,n) \in \mathbf{Z}^2$ has dimension $c(mn)$ whenever $(m,n) \neq (0,0)$. The monster should be thought of as a group of "diagram automorphisms" of this Lie algebra, in the same way that the symmetric group S_3 is a group of diagram automorphisms of the Lie algebra D_4. The monster Lie algebra has a denominator formula, similar to the Weyl denominator formula for finite dimensional Lie algebras and the Macdonald-Kac identities for affine Lie algebras, which looks like

$$j(\sigma) - j(\tau) = p^{-1} \prod_{\substack{m>0 \\ n \in \mathbf{Z}}} (1 - p^m q^n)^{c(mn)}$$

where $p = e^{2\pi i \sigma}$, $q = e^{2\pi i \tau}$. This formula was discovered independently in the 1980's by several people, including Koike, Norton, and Zagier. There are similar identities with $j(\tau)$ replaced by the McKay-Thompson series of any element of the monster, and Cummins and Gannon showed that any function satisfying such identities is a Hauptmodul. So this provides some sort of explanation of Conway and Norton's observation that the McKay-Thompson series are all Hauptmoduls.

So the question "What is the monster" now has several reasonable answers:

1 The monster is the largest sporadic simple group, or alternatively the unique simple group with its order.

2 It is the automorphism group of the Griess algebra.

3 It is the automorphism group of the monster vertex algebra. (This is probably the best answer.)

4 It is a group of diagram automorphisms of the monster Lie algebra.

Unfortunately none of these definitions is completely satisfactory. At the moment all constructions of the algebraic structures above seem artificial; they are constructed as sums of two or more apparently unrelated spaces, and it takes a lot of effort to define the algebraic structure on the sum of these spaces and to check that the monster acts on the resulting structure.

References.

[A] J. H. Conway, R. T. Curtis, S. P. Norton, R. A. Parker, R. A. Wilson, Atlas of finite groups, Clarendon Press, Oxford, 1985.

[C-N] J. H. Conway, S. Norton, Monstrous moonshine, Bull. London. Math. Soc. 11 (1979) 308-339.

[K] V. Kac, Vertex algebras for beginners. Second edition. University Lecture Series, 10. American Mathematical Society, Providence, RI, 1998.

arXiv:hep-th/9408037v1 5 Aug 1994

Realizations of the Monster Lie Algebra

Elizabeth Jurisich*
James Lepowsky
Robert L. Wilson

Department of Mathematics, Rutgers University
New Brunswick, NJ 08903
*Current address: Department of Mathematics, University of Chicago
Chicago, IL 60637

e-mail addresses: jurisich@math.rutgers.edu, lepowsky@math.rutgers.edu, rwilson@math.rutgers.edu

1 Introduction

In this paper we reinterpret the theory of generalized Kac-Moody Lie algebras in terms of local Lie algebras formed from reductive or Kac-Moody algebras and certain modules (possibly infinite-dimensional) for these algebras. We exploit the fact, established in [19], that certain generalized Kac-Moody algebras contain specific large free subalgebras, in order to exhibit these generalized Kac-Moody algebras as explicitly prescribed Lie algebras of operators acting on tensor algebras. In the most important special case, that of R. Borcherds' Monster Lie algebra $\mathfrak{m}$, introduced in [4] (see also [5]), we apply the free subalgebra result [19] to simplify Borcherds' work [4] on the Conway-Norton conjectures (see [8]) for the "moonshine module" $V^\natural$ ([11], [12]) for the Fischer-Griess Monster group M. In particular, we realize $\mathfrak{m}$ as an explicitly prescribed M-covariant Lie algebra of operators acting on the tensor algebra over a certain $\mathfrak{gl}_2$- and M-module built in a simple way from the M-module $V^\natural$.

In [4], Borcherds produces recursion relations which, along with initial conditions, uniquely characterize the McKay-Thompson series (i.e., the graded traces) of the elements of M acting on the infinite-dimensional M-module $V^\natural$ constructed in [11], [12]. In this way he completes the proof of the Conway-Norton conjectures as they relate to $V^\natural$, in the sense that he shows that the McKay-Thompson series for $V^\natural$ do indeed agree with the modular functions listed in [8] for all elements of M, since the coefficients of those modular functions satisfy the same recursion relations (replication formulas) and initial conditions. For instance, for the identity element of M the corresponding McKay-Thompson

series is the modular function $J(q) = j(q) - 744$, which is the graded dimension of $V^\natural$. (This is one of the McKay-Thompson series for $V^\natural$ already determined in [11], [12].) The proof in [4] involves defining the Lie algebra $\mathfrak{m}$, establishing a triangular decomposition $\mathfrak{m} = \mathfrak{n}^+ \oplus \mathfrak{h} \oplus \mathfrak{n}^-$ and a denominator identity for $\mathfrak{m}$, and computing the homology of the subalgebra $\mathfrak{n}^+ \subset \mathfrak{m}$ (or equivalently, of the subalgebra $\mathfrak{n}^-$). Borcherds is then able to show that the coefficients of the McKay-Thompson series satisfy certain recursion relations. As we will explain below, much of the theory involved in the proof in [4] is illuminated if we decompose $\mathfrak{m}$ as

$$\mathfrak{m} = \mathfrak{u}^+ \oplus \mathfrak{gl}_2 \oplus \mathfrak{u}^-,$$

where $\mathfrak{u}^+$ and $\mathfrak{u}^-$ are the Lie algebras proved to be free in [19], and if we consider the $\mathfrak{gl}_2$-module $\mathfrak{u}^-$ in place of $\mathfrak{n}^-$. We base our considerations on this decomposition rather than the triangular decomposition used in [4].

The properties of the Monster Lie algebra are of central importance to the proof in [4]. In Section 2 we recall facts about generalized Kac-Moody algebras, of which $\mathfrak{m}$ is an example. Borcherds' original work on the subject [2] contains results such as character formulas and a denominator identity for generalized Kac-Moody algebras. V. Kac in [21] gives an outline (without detail) of how to rigorously develop the theory of generalized Kac-Moody algebras by indicating that one should follow the arguments presented there for Kac-Moody algebras; see also [17]. Included in [18] is an exposition developing the theory of generalized Kac-Moody algebras, including an extension of the homology results of [14] (not covered in [21]) to these more general Lie algebras. That this can be done is mentioned and used in [4]. This homology theory gives another proof of the character and denominator formulas (see [18]). We find it appropriate to work with the extended Lie algebra as in [14] and [24] (that is, the Lie algebra with suitable degree derivations adjoined). Equivalently, one can generalize the theorems in [21]. In either of these approaches the Cartan subalgebra is sufficiently enlarged to make the simple roots linearly independent and of multiplicity one.

It is shown in [19] that any generalized Kac-Moody algebra $\mathfrak{g}$ that has no mutually orthogonal imaginary simple roots can be decomposed as $\mathfrak{g} = \mathfrak{u}^+ \oplus (\mathfrak{g}_S + \mathfrak{h}) \oplus \mathfrak{u}^-$ where $\mathfrak{g}_S$ is a (semisimple or) Kac-Moody Lie algebra and $\mathfrak{u}^+$ and $\mathfrak{u}^-$ are free Lie algebras over particular modules for $\mathfrak{g}_S$. This theorem is recalled in Section 2 below, where we also summarize properties of $\mathfrak{m}$ which we use. It is also shown that there are Lie algebras which cover $\mathfrak{m}$ such that the natural action of M on $\mathfrak{m}$ lifts to an action of M on these larger generalized Kac-Moody algebras.

Facts about local Lie algebras that we use are also included in Section 2. Mainly, we recall that the maximal Lie algebra $\mathfrak{a}_{max}$ associated to a local Lie algebra $\mathfrak{a} = \mathfrak{a}_1 \oplus \mathfrak{a}_0 \oplus \mathfrak{a}_{-1}$ is such that $\mathfrak{a}_{max}^{\pm}$ is the free Lie algebra on $\mathfrak{a}_{\pm 1}$, as in [20]. In light of the results of [19] mentioned above, it is not surprising that these generalized Kac-Moody algebras with suitable derivations adjoined are closely related to maximal Lie algebras $\mathfrak{a}_{max}$ associated to the appropriate

local Lie algebras $\mathfrak{a}$. We show this in Section 3.

In Section 4, we discuss some examples of standard (irreducible) modules which have the special property that they are induced from irreducible modules of the subalgebra $(\mathfrak{g}_S + \mathfrak{h}) \oplus \mathfrak{u}^+$. As mentioned above, combined with the free subalgebra result, this gives yet another natural realization of the Monster Lie algebra, this time as an explicitly prescribed M-covariant Lie algebra of operators on the space $T(\mathcal{V})$, where $\mathcal{V}$ is one of the two (infinite-dimensional) $\mathfrak{gl}_2$- and M-modules $\mathfrak{a}_{\pm 1}$ occurring in the local Lie algebra construction (here $\mathfrak{a}_0 = \mathfrak{gl}_2$). The examples of standard modules that we consider in this paper are all "generalized Verma modules," that is, they are induced from irreducible modules for "parabolic subalgebras" (e.g., the subalgebra $\mathfrak{p} = (\mathfrak{g}_S + \mathfrak{h}) \oplus \mathfrak{u}^+$ for a generalized Kac-Moody algebra $\mathfrak{g} = \mathfrak{u}^- \oplus (\mathfrak{g}_S + \mathfrak{h}) \oplus \mathfrak{u}^+$). The considerations on generalized Verma modules in [14] and [25] remain applicable to the case of generalized Kac-Moody algebras. If we choose the right parabolic subalgebra and highest weight λ, then the corresponding generalized Verma module is a standard module, and in particular, it is irreducible. In the case where $\mathfrak{u}^-$ is a free Lie algebra over a vector space V, the tensor algebra over V clearly has the structure of a generalized Verma module (see Section 4). As we have mentioned, this includes the case of the Monster Lie algebra $\mathfrak{m}$. The "fundamental modules" constructed in [15] for those generalized Kac-Moody algebras (or Borcherds algebras, as they are called in [15]) with one simple imaginary root are special cases of these induced modules. For the case of $\mathfrak{m}$, we conjecture that the weight-zero components v_0 of the vertex operators $Y(v, z)$ for $v \in \mathcal{V}$ generate a free associative algebra of operators on $\mathfrak{m}$. This may provide a more conceptual framework for studying the irreducible $\mathfrak{m}$-modules $T(\mathcal{V})$. (See Sections 2 and 4.)

Finally, in Section 5, we use the free Lie algebra structure of $\mathfrak{u}^-$ to simplify part of the proof appearing in [4] concerning the replication formulas. We follow the general ideas of Borcherds' proof, but we use the subalgebra $\mathfrak{u}^-$ in place of $\mathfrak{n}^-$. Because of its structure as a free Lie algebra, computations involving $\mathfrak{u}^-$ are comparatively simple. For example, the homology of $\mathfrak{u}^-$ with coefficients in the trivial module is elementary to compute, while providing us with essentially the same information as the homology of $\mathfrak{n}^-$. Moreover, the formula, obtained from the Euler-Poincaré identity, that leads to replication (now symmetric in the formal variables p and q) is visibly similar to formulas occurring in such classical references on free Lie algebras as [6] and [28] (as is the proof of the denominator identities in [19]). We are able to obtain the replication formula as it appears in [27] by carrying out an analog of a computation in [6]. After proving a Möbius inversion formula consistent with Adams operations, we establish recursion relations equivalent to replication. Even though replication is more easily understood in this context, it is still not clear why series satisfying these replication formulas should automatically have modular transformation properties. Philosophically, the fact that $\mathfrak{n}^+$ (or $\mathfrak{n}^-$) contains such a large free Lie algebra explains why computations such as determining root multiplicities

and the replication formulas are at all manageable for $\mathfrak{m}$.

This paper was presented by E. J. and J. L. at the June, 1994 AMS-IMS-SIAM Joint Summer Research Conference on Moonshine, the Monster and Related Topics at Mount Holyoke College.

We are grateful to S.-J. Kang for informing us that formula (37) in this paper has been independently obtained in joint work with V. Kac.

We would like to thank R. Borcherds, F. Knop and C. Weibel for helpful comments. We also thank R. Gebert and J. Teschner, K. Harada, M. Miyamoto and H. Yamada, S. J. Kang and S. Naito for sending us preprints of their works on generalized Kac-Moody algebras. E. J. is supported in part by a Rutgers University Louis Bevier Graduate Fellowship, and J. L. and R. W. are supported in part by NSF grant DMS-9111945.

2 Preliminary Material

2.1 Notation

For any vector space V we let $L(V)$ denote the free Lie algebra over V, and $T(V)$ the tensor algebra over V. Recall that $T(V)$ is the universal enveloping algebra of $L(V)$. Let V^* denote the dual space of V. For a $\mathbb{Z}$-graded vector space $V = \coprod_{i\in\mathbb{Z}} V_i$ we let $V^{\pm}$ denote $\coprod_{i\in\pm\mathbb{Z}_+} V_i$. Unless otherwise indicated, vector spaces will be over $\mathbb{C}$. We denote the Monster group by M and the "moonshine module" for M by $V^\natural$, as in [12].

2.2 Generalized Kac-Moody algebras

The theory of generalized Kac-Moody Lie algebras with imaginary simple roots was developed by Borcherds starting in [2]. Two further definitions (different from the first and from one another) are given in [3] and [4]. Of these definitions the one given below most resembles the one appearing in [2], where there is a definition in terms of an index set I, a vector space H, a set of generators $H \cup \{e_i, f_i |\ i \in I\}$, and relations, determined by square matrices, which agree with the ones given here. Borcherds does not divide out by the radical (that is, the largest graded ideal having trivial intersection with $\mathfrak{h}$) in his definition of generalized Kac-Moody algebra. In [3] Borcherds extends his definition to include some additional elements h_{ij} which are central. These elements do not enter into our discussion. In [4] Borcherds defines generalized Kac-Moody algebras in terms of "an almost positive definite bilinear form" and an involution, and it is a theorem stated in [4] (see also [17] and [19] for clarification of the statement and for two different detailed proofs; cf. [21]) that such Lie algebras are closely related to Lie algebras defined from matrices.

The definition given here of the generalized Kac-Moody algebra $\mathfrak{g}(A)$ (see below) is the same as that in [19]. In [18] and [21] a generalized Kac-Moody

algebra is defined as $\mathfrak{g}(A)$ modulo its radical. The radical of $\mathfrak{g}(A)$ is in fact zero (see [17] and [18] for detailed proofs, clarifying [2], using [13] or [21]; or see [19], where there is a relatively elementary argument for the case of Lie algebras satisfying the hypothesis of Theorem 2.1 below—which are the main examples discussed in this paper—showing that the vanishing of the radical for Kac-Moody algebras implies the vanishing of the radical for this class). Thus the definition given here is equivalent to those in [18] and [21].

Theorem 2.1, given at the end of this section, is the main theorem of [19], and is used throughout this paper.

We will use the notation of [14], [24] and [25], suitably generalized to the case of generalized Kac-Moody Lie algebras in [18]. We restrict our attention to the case of symmetric matrices (although symmetrizable matrices could be easily handled) and countable index sets.

Let I be a (finite or) countable index set and let $A = (a_{ij})_{i,j\in I}$ be a matrix with entries in $\mathbb{R}$, satisfying the following conditions:

(C1) A is symmetric.

(C2) If $i \neq j$ $(i, j \in I)$, then $a_{ij} \leq 0$.

(C3) If $a_{ii} > 0$ $(i \in I)$, then $\frac{2a_{ij}}{a_{ii}} \in \mathbb{Z}$ for all $j \in I$.

Let $\mathfrak{g}(A)$ be the Lie algebra with generators h_i, e_i, f_i, $i \in I$, and the following defining relations: For all $i, j, k \in I$,

(R1) $[h_i, h_j] = 0$,

(R2) $[h_i, e_k] - a_{ik}e_k = 0$,

(R3) $[h_i, f_k] + a_{ik}f_k = 0$,

(R4) $[e_i, f_j] - \delta_{ij}h_i = 0$,

(R5) $(\operatorname{ad} e_i)^{\frac{-2a_{ij}}{a_{ii}}+1} e_j = 0$ and $(\operatorname{ad} f_i)^{\frac{-2a_{ij}}{a_{ii}}+1} f_j = 0$ for all $i \neq j$ with $a_{ii} > 0$,

(R6) $[e_i, e_j] = 0$ and $[f_i, f_j] = 0$ whenever $a_{ij} = 0$.

Definition: The Lie algebra $\mathfrak{g}(A)$ is *the generalized Kac-Moody (Lie) algebra associated to the matrix A*. Any Lie algebra of the form $\mathfrak{g}(A)/\mathfrak{c}$ where $\mathfrak{c}$ is a central ideal is called a *generalized Kac-Moody algebra*.

In $\mathfrak{g}(A)$, the elements h_i, e_i, f_i for $i \in I$ are linearly independent. Define the *Cartan subalgebra* $\mathfrak{h}$ of $\mathfrak{g}(A)$ to be the span of the h_i for $i \in I$. The center of $\mathfrak{g}(A)$ lies in $\mathfrak{h}$.

Let A satisfy $(C1) - (C3)$ and let $S \subset I$ be the set of all indices i such that $a_{ii} > 0$. (Many of the following considerations apply for more general subsets of I; it is for convenience that we fix S as indicated.) Let B be the the submatrix of A given by $(a_{ij})_{i,j\in S}$.

The associated Lie algebra $\mathfrak{g}(B)$ is a Kac-Moody algebra (with a symmetric defining matrix). Let $\mathfrak{g}_S$ be the Lie subalgebra of $\mathfrak{g}(A)$ generated by $\{e_i, h_i, f_i\}_{i\in S}$. Then $\mathfrak{g}_S$ is isomorphic to $\mathfrak{g}(B)$. Note that we are using the characterization of $\mathfrak{g}(B)$ in terms of generators and relations, i.e., we are using the fact that the radical is zero (see [13] or [21]; this is Serre's theorem [29] in case B is of finite type). For the important case of the Monster Lie algebra $\mathfrak{m}$, described in more detail in the next section, $B = (2)$ and $\mathfrak{g}_S \cong \mathfrak{sl}_2$.

The Lie algebra $\mathfrak{g} = \mathfrak{g}(A)$ is naturally graded by $\mathbb{Z}^I$. Let D_i be the ith degree derivation of $\mathfrak{g}$ with respect to this grading, i.e., $D_i e_j = \delta_{ij} e_j$, $D_i f_j = -\delta_{ij} f_j$ and $D_i h_j = 0$ for all $i, j \in I$. The span of the D_i for $i \in I$ forms an abelian Lie algebra $\mathfrak{d}_0$ of derivations of $\mathfrak{g}$. Given a subspace $\mathfrak{d}$ of $\mathfrak{d}_0$ let $\mathfrak{g}^e$ be the semidirect product $\mathfrak{d} \ltimes \mathfrak{g}$. Denote by $\mathfrak{h}^e$ the abelian subalgebra $\mathfrak{d} \oplus \mathfrak{h}$. The Lie algebra $\mathfrak{g}^e$ is called the *extended* Lie algebra. Call $\mathfrak{h}^e$ the *Cartan subalgebra* of $\mathfrak{g}^e$.

Define the *simple roots* $\alpha_i \in (\mathfrak{h}^e)^*$ for $i \in I$ by the conditions $[h, e_i] = \alpha_i(h) e_i$ for all $h \in (\mathfrak{h}^e)^*$, $i \in I$. Fix a subspace $\mathfrak{d}$ of $\mathfrak{d}_0$ such that the simple roots α_i are linearly independent. There is a unique symmetric bilinear form $(\cdot, \cdot)$ on the span of the α_i, $i \in I$, determined by the conditions $(\alpha_i, \alpha_j) = a_{ij}$. This can be extended (not necessarily uniquely) to a symmetric bilinear form, which we continue to call $(\cdot, \cdot)$, on all of $(\mathfrak{h}^e)^*$ such that $(\varphi, \alpha_i) = \varphi(h_i)$ for all $\varphi \in (\mathfrak{h}^e)^*$.

For $\alpha \in (\mathfrak{h}^e)^*$ define

$$\mathfrak{g}^\alpha = \{x \in \mathfrak{g} | [h, x] = \alpha(h)x \text{ for all } h \in \mathfrak{h}^e\}.$$

Nonzero elements $\alpha \in (\mathfrak{h}^e)^*$ such that $\mathfrak{g}^\alpha \neq 0$ are called *roots*. Let $\Delta \subset (\mathfrak{h}^e)^*$ be the set of roots, and Δ_+ the set of roots which are nonnegative integral linear combinations of α_i's. The roots in Δ_+ are called *positive* roots. Let $\Delta_- = -\Delta_+$, the set of *negative* roots. All of the roots are either positive or negative. There is a root space decomposition

$$\mathfrak{g} = \coprod_{\varphi\in\Delta_+} \mathfrak{g}^\varphi \oplus \mathfrak{h} \oplus \coprod_{\varphi\in\Delta_-} \mathfrak{g}^\varphi. \tag{1}$$

Define the set of *real* roots Δ_R to be the set of $\alpha \in \Delta$ such that $(\alpha, \alpha) > 0$. The set of *imaginary* roots is $\Delta \backslash \Delta_R$. The imaginary simple roots are those α_i for which $a_{ii} \leq 0$.

The symmetric bilinear form on the span of the α_i, $i \in I$, can be transferred to $\mathfrak{h}$ by letting $(h_i, h_j)_\mathfrak{h} = (\alpha_i, \alpha_j) = a_{ij}$. This form can be extended (not uniquely, for it depends upon a choice of symmetric bilinear form on $\mathfrak{d}$) to a symmetric bilinear form $(\cdot, \cdot)_{\mathfrak{h}^e}$ on $\mathfrak{h}^e = \mathfrak{d} \oplus \mathfrak{h}$, satisfying the conditions $(d, h_j) = \alpha_j(d)$ for $d \in \mathfrak{d}$ and $j \in I$. We assume that this form on $\mathfrak{h}^e$ is consistent with the form on $(\mathfrak{h}^e)^*$. There is a unique invariant symmetric bilinear form $(\cdot, \cdot)_{\mathfrak{g}^e}$ on $\mathfrak{g}^e$ which extends $(\cdot, \cdot)_{\mathfrak{h}^e}$. This form satisfies the condition

$$(\mathfrak{g}^\varphi, \mathfrak{g}^\psi)_{\mathfrak{g}^e} = (\mathfrak{h}^e, g^\varphi)_{\mathfrak{g}^e} = 0$$

for all $\varphi, \psi \in \Delta$ such that $\varphi + \psi \neq 0$. Given any $\varphi = \sum_{i \in I} n_i \alpha_i \in \Delta$ ($n_i \in \mathbb{Z}$), we define $h_\varphi = \sum_{i \in I} n_i h_i$. Then for $a \in \mathfrak{g}^\varphi$, $b \in \mathfrak{g}^{-\varphi}$, the form $(\cdot, \cdot)_{\mathfrak{g}^e}$ satisfies the condition

$$[a, b] = (a, b) h_\varphi.$$

In particular, $(e_i, f_j)_{\mathfrak{g}^e} = \delta_{ij}$ for $i, j \in I$. (See [18] for details; cf. [20], [24], [26].)

We assume that in the case in which A is a (finite) generalized Cartan matrix, the space $\mathfrak{d}$ and the form on $\mathfrak{d}$ are chosen so that the form $(\cdot, \cdot)_{\mathfrak{h}^e}$ on the (finite-dimensional) space $\mathfrak{h}^e$ is nonsingular, so that the Lie algebra $\mathfrak{g}^e = \mathfrak{g}(A)^e$ is, in this case, the usual Kac-Moody algebra (cf. [21]) and the form $(\cdot, \cdot)_{\mathfrak{g}^e}$ is nondegenerate.

The center of the unextended Lie algebra $\mathfrak{g}$ is equal to the radical of the form $(\cdot, \cdot)$ restricted to $\mathfrak{g}$. We also note that any element of the center $h = \sum_{j \in I} c_j h_j$ ($c_j \in \mathbb{C}$, $c_j = 0$ for almost all $j \in I$) provides a linear dependence relation $\sum_{j \in I} c_j a_{ij}$, among the columns of the matrix A and conversely, any such relation among columns of A determines an element of the center of $\mathfrak{g}$.

We make the following definitions: $\Delta^S = \Delta \cap \coprod_{i \in S} \mathbb{Z}\alpha_i$, $\Delta_+^S = \Delta_+ \cap \Delta^S$ and $\Delta_-^S = \Delta_- \cap \Delta^S$. Denote by $\mathfrak{h}_S$ the span of the h_i, $i \in S$. Then we have the root space decomposition:

$$\mathfrak{g}_S = \coprod_{\varphi \in \Delta_+^S} \mathfrak{g}^\varphi \oplus \mathfrak{h}_S \oplus \coprod_{\varphi \in \Delta_-^S} \mathfrak{g}^\varphi. \tag{2}$$

Define the following subalgebras of $\mathfrak{g} = \mathfrak{g}(A)$:

$$\mathfrak{n}^+ = \coprod_{\varphi \in \Delta_+} \mathfrak{g}^\varphi; \quad \mathfrak{n}^- = \coprod_{\varphi \in \Delta_-} \mathfrak{g}^\varphi; \quad \mathfrak{n}_S^+ = \coprod_{\varphi \in \Delta_+^S} \mathfrak{g}^\varphi;$$

$$\mathfrak{n}_S^- = \coprod_{\varphi \in \Delta_-^S} \mathfrak{g}^\varphi; \quad \mathfrak{u}^+ = \coprod_{\varphi \in \Delta_+ \setminus \Delta_+^S} \mathfrak{g}^\varphi; \quad \mathfrak{u}^- = \coprod_{\varphi \in \Delta_- \setminus \Delta_-^S} \mathfrak{g}^\varphi; \quad \mathfrak{r} = \mathfrak{g}_S + \mathfrak{h}. \tag{3}$$

Let $\mathfrak{p} = \mathfrak{r} \oplus \mathfrak{u}^+$, the "parabolic subalgebra" of $\mathfrak{g}$ defined by S. Note that

$$\mathfrak{g} = \mathfrak{u}^- \oplus \mathfrak{p}.$$

The subalgebras $\mathfrak{p}^e$, $\mathfrak{r}^e$ and $\mathfrak{g}_S^e$ are defined in the obvious ways.

Let X be an $\mathfrak{h}^e$-module, e.g., a $\mathfrak{g}^e$-module considered as an $\mathfrak{h}^e$-module by restriction, and let $\nu \in (\mathfrak{h}^e)^*$. Define the corresponding *weight space*

$$X_\nu = \{x \in X | h \cdot x = \nu(h)x \text{ for all } h \in \mathfrak{h}^e\}.$$

We say that ν is a *weight* if $X_\nu \neq 0$ and that the nonzero elements of X_ν are *weight vectors* of weight ν. The module X is called a *weight module* if X is a direct sum of its weight spaces.

A $\mathfrak{g}^e$-module is called a *highest weight module* if it is generated by a weight vector x annihilated by $\mathfrak{n}^+$. The weight of x is called the *highest weight* of

X. For any highest weight module X we have $U(\mathfrak{n}^-) \cdot x = X$. Lowest weight modules are defined analogously.

Concepts defined above for $\mathfrak{g}(A)$ such as extended Lie algebra, roots, simple roots, Cartan subalgebra and weights can be defined analogously for a generalized Kac-Moody algebra of the form $\mathfrak{g}(A)/\mathfrak{c}$ with $\mathfrak{c}$ a subspace of the center. Note that in particular, the roots of $\mathfrak{g}(A)$, which are elements of $(\mathfrak{h}^e)^*$, vanish on $\mathfrak{c} \subset \mathfrak{h}$ and may therefore be identified with the roots of $\mathfrak{g}(A)/\mathfrak{c}$, which are elements of the dual of $(\mathfrak{h}/\mathfrak{c})^e = \mathfrak{d} \oplus \mathfrak{h}/\mathfrak{c}$. We shall thereby identify the roots, simple roots, the e_i's and the f_i's (but not the h_i's), etc., of $\mathfrak{g}(A)$ with those of $\mathfrak{g}(A)/\mathfrak{c}$. The simple roots remain linearly independent. For those $\mathfrak{g}(A)^e$ weight modules X on which $\mathfrak{c}$ acts trivially, X is also a $(\mathfrak{g}(A)/\mathfrak{c})^e$-module, and we identify the weights, etc., for X as $\mathfrak{g}(A)^e$ and $(\mathfrak{g}(A)/\mathfrak{c})^e$-modules.

We shall use the following theorem from [19]:

Theorem 2.1 *Let $\mathfrak{g}(A) = \mathfrak{g}$ be a generalized Kac-Moody algebra such that if α_i and α_j are two distinct imaginary simple roots then $a_{ij} < 0$. Let $S = \{i \in I | a_{ii} > 0\}$. Then $\mathfrak{g} = \mathfrak{u}^+ \oplus \mathfrak{r} \oplus \mathfrak{u}^-$, where $\mathfrak{u}^-$ is the free Lie algebra on the direct sum of the integrable highest weight $\mathfrak{g}_S$-modules $U(\mathfrak{n}_S^-) \cdot f_j$ for $j \in I \backslash S$ and $\mathfrak{u}^+$ is the free Lie algebra on the direct sum of the integrable lowest weight $\mathfrak{g}_S$-modules $U(\mathfrak{n}_S^+) \cdot e_j$ for $j \in I \backslash S$.*

2.3 The Monster Lie algebra

The Monster Lie algebra $\mathfrak{m}$ is defined by Borcherds in [4] (see [19] for more detail and further references). It is constructed using the vertex operator algebra $V^\natural$ and the vertex algebra V_L associated to an even Lorentzian lattice L of rank 2, which we identify with $\mathbb{Z} \oplus \mathbb{Z}$, equipped with the bilinear form $\langle \cdot, \cdot \rangle$ given by the matrix $\begin{pmatrix} 0 & -1 \\ -1 & 0 \end{pmatrix}$. The Lie algebra $\mathfrak{m}$ is a certain quotient of the "physical subspace" P_1 of the vertex algebra $V^\natural \otimes V_L$ (P_1 is the space of lowest weight vectors of weight one for the Virasoro algebra): $\mathfrak{m} = P_1/R$, where R is the radical of a natural symmetric bilinear form. We do not need the details of the actual definition in this paper, but will quote results from [4], [12] and [19] as they are needed. (We note that Borcherds' construction of the Lie algebra is over $\mathbb{R}$, while we are working over $\mathbb{C}$.)

The vertex operator algebra $V^\natural$ is equipped with an action of M, and $V^\natural$ has a natural M-invariant symmetric nondegenerate bilinear form (see [12]).

We grade $V^\natural$ as $\coprod_{i \geq -1} V_i^\natural$ where $V_i^\natural$ is the span of all vectors of conformal weight $i+1$. Recall ([11], [12]) that $V^\natural$ has graded dimension $\sum_{i \geq -1} \left(\dim V_i^\natural \right) q^i = J(q)$, where $J(q) = \sum_{i \geq -1} c(i) q^i$ is the modular function $j(q) - 744$, so that $\dim V_i^\natural = c(i)$ for all $i \geq -1$. We have $c(-1) = 1,\ c(0) = 0,\ c(1) = 196884$. (For $i < -1$ we set $V_i^\natural = 0$ and $c(i) = 0$.) The action of M preserves each $V_i^\natural$.

The rank 2 lattice L provides a $\mathbb{Z} \oplus \mathbb{Z}$-grading of the Lie algebra $\mathfrak{m}$. The group action of M on the vertex operator algebra $V^\natural$ gives an action of M on the vertex algebra $V^\natural \otimes V_L$, where M acts trivially on V_L. This action induces an action of M as Lie algebra automorphisms on $\mathfrak{m}$, preserving the grading. The Lie algebra $\mathfrak{m}$ has a natural nondegenerate invariant symmetric bilinear form which is also M-invariant. This form satisfies the condition $(\mathfrak{m}_r, \mathfrak{m}_s) = 0$ unless $r + s = 0$, for $r, s \in \mathbb{Z} \oplus \mathbb{Z}$. Lemma 2.2 below is one of the main theorems of [4] (cf. [19] for some clarification), and follows from the "no-ghost" theorem from string theory. Although the statement of the theorem [16] involves vector spaces over $\mathbb{R}$, after the linear isomorphism given below is established over $\mathbb{R}$ we extend the isomorphism to $\mathbb{C}$.

Lemma 2.2 *As a $\mathbb{Z} \oplus \mathbb{Z}$-graded Lie algebra and M-module, $\mathfrak{m}$ satisfies:*

$$\mathfrak{m} = \left(\coprod_{m,n<0} \mathfrak{m}_{(m,n)} \right) \oplus \mathfrak{m}_{(-1,1)} \oplus \mathfrak{m}_{(0,0)} \oplus \mathfrak{m}_{(1,-1)} \oplus \left(\coprod_{m,n>0} \mathfrak{m}_{(m,n)} \right), \qquad (4)$$

where there are natural isomorphisms

$$\mathfrak{m}_{(m,n)} \cong V^\natural_{mn} \text{ as an } M\text{-module for } (m,n) \neq (0,0),$$

$$\mathfrak{m}_{(0,0)} \cong \mathbb{C} \oplus \mathbb{C}, \text{ a trivial } M\text{-module.}$$

Furthermore, the M-invariant bilinear form is preserved under these isomorphisms, i.e., the pairing between $\mathfrak{m}_{(m,n)}$ and $\mathfrak{m}_{(-m,-n)}$ corresponds to the bilinear form on $V^\natural_{mn}$. $\square$

We define

$$\mathfrak{n}^+ = \coprod_{m,n>0} \mathfrak{m}_{(m,n)} \oplus \mathfrak{m}_{(1,-1)},$$

and

$$\mathfrak{n}^- = \coprod_{m,n<0} \mathfrak{m}_{(m,n)} \oplus \mathfrak{m}_{(-1,1)},$$

so that

$$\mathfrak{m} = \mathfrak{n}^- \oplus \mathfrak{m}_{(0,0)} \oplus \mathfrak{n}^+.$$

It follows from the definition of $\mathfrak{m}$ that

$$\mathfrak{m}_{(-1,1)} \oplus \mathfrak{m}_{(0,0)} \oplus \mathfrak{m}_{(1,-1)} \cong \mathfrak{gl}_2.$$

To make this explicit, let

$$1 \to \langle \kappa \rangle \to \hat{L} \bar{\to} L \to 1$$

be the central extension of L with κ of order 2 and commutator map given by $\kappa^{\langle \alpha, \beta \rangle}$, $\alpha, \beta \in L$ (cf. [12, Ch. 7]). We choose elements $a, b \in \hat{L}$ to satisfy

$\bar{a} = (1,1)$, $\bar{b} = (1,-1) \in L$. As in [19], appropriate elements of $V^\natural \otimes V_L$ are used to denote their equivalence classes in $\mathfrak{m}$. Let ι denote the natural map from $\hat{L}$ to V_L, so that for $c \in \hat{L}$, $\iota(c)$ is identified with $1 \otimes \iota(c) \in 1 \otimes \mathbb{C}\{L\} \subset S \otimes \mathbb{C}\{L\} = V_L$, where S is the symmetric algebra of the negatively graded subspace of the appropriate Heisenberg algebra and where $\mathbb{C}\{L\}$ is the "twisted group algebra" of L (cf. [12, Ch. 7]).

Consider the elements $e = 1 \otimes \iota(b) \in \mathfrak{m}_{(1,-1)}$, $f = 1 \otimes \iota(b^{-1}) \in \mathfrak{m}_{(-1,1)}$, $h = \bar{b}(-1) \otimes \iota(1) \in \mathfrak{m}_{(0,0)}$ and $z = \bar{a}(-1) \otimes \iota(1) \in \mathfrak{m}_{(0,0)}$ in $\mathfrak{m}$. Then e, f and h span a copy of $\mathfrak{sl}_2$, with relations $[e,f] = h$, $[h,e] = 2e$ and $[h,f] = -2f$. Since z commutes with e, f and h, the above elements span a copy of $\mathfrak{gl}_2$ in $\mathfrak{m}$. Under the adjoint action $\mathfrak{m}$ is a $\mathfrak{gl}_2$-module, and the element z acts on $\mathfrak{m}_{(m,n)}$ as the scalar $-m-n$ and h acts on $\mathfrak{m}_{(m,n)}$ as the scalar $m-n$. Note that $\mathfrak{m}_{(-1,1)} \oplus \mathfrak{m}_{(0,0)} \oplus \mathfrak{m}_{(1,-1)}$ is a trivial M-module. Consequently, $\mathfrak{m}$ is a $(\mathfrak{gl}_2, M)$-module, i.e., $\mathfrak{m}$ is both a $\mathfrak{gl}_2$- and an M-module and these actions commute.

The Lie algebra $\mathfrak{m}$ is a generalized Kac-Moody algebra (see [4], where the definition used there allows that $\mathfrak{m}$ might be a generalized Kac-Moody algebra with some possibly nontrivial "outer" derivations adjoined, and see also [19]). In fact, $\mathfrak{m}$ is shown in [19] to be isomorphic to $\mathfrak{g}(A)/\mathfrak{c}$, where $\mathfrak{g}(A)$ is the generalized Kac-Moody algebra associated to the following matrix A and $\mathfrak{c}$ is the full center of $\mathfrak{g}(A)$:

$$A = \left(\begin{array}{c|ccc|ccc|c} 2 & 0 & \cdots & 0 & -1 & \cdots & -1 & \cdots \\ \hline 0 & -2 & \cdots & -2 & -3 & \cdots & -3 & \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \cdots \\ 0 & -2 & \cdots & -2 & -3 & \cdots & -3 & \\ \hline -1 & -3 & \cdots & -3 & -4 & \cdots & -4 & \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \cdots \\ -1 & -3 & \cdots & -3 & -4 & \cdots & -4 & \\ \hline \vdots & & \vdots & & & \vdots & & \end{array}\right).$$

Consider the index set $\mathcal{I}_0 = \{-1, 1, 2, 3, \cdots\}$. We shall take our index set I, indexing the generators and simple roots of $\mathfrak{g}(A)$, to be the set $\mathcal{I}$ of pairs $(i,k) \in \mathcal{I}_0 \oplus \mathbb{Z}_+$ with $1 \le k \le c(i)$, reflecting the block form of the matrix. Thus the block in position $(i,j) \in \mathcal{I}_0 \times \mathcal{I}_0$ has size $c(i) \times c(j)$, and entries $-(i+j)$ for $i, j \in \mathcal{I}_0$. Then $\mathfrak{g}(A)$ has generators e_{ik}, h_{ik}, f_{ik} and simple roots α_{ik}, $(i,k) \in \mathcal{I}$. We will write $e_{-1,1}$ as e_{-1}, $f_{-1,1}$ as f_{-1} and $h_{-1,1}$ as h_{-1}. Note that the simple roots α_{ik} for fixed $i \in \mathcal{I}_0$, $1 \le k \le c(i)$, all have the same values on the Cartan subalgebra $\mathfrak{h} \subset \mathfrak{h}^e$. Under the isomorphism between $\mathfrak{g}(A)/\mathfrak{c}$ and $\mathfrak{m}$, the subalgebras $\mathfrak{n}^\pm \subset \mathfrak{g}(A)$ are identified with the subalgebras $\mathfrak{n}^\pm \subset \mathfrak{m}$ and the elements h_{ik} for $(i,k) \in \mathcal{I}$ map to $\left(\frac{1-i}{2}\right) h + \left(\frac{1+i}{2}\right) z \in \mathfrak{m}_{(0,0)}$. The $\mathbb{Z} \oplus \mathbb{Z}$-grading of the Lie algebra $\mathfrak{m}$ given by (4) is determined by the root space grading of $\mathfrak{m}$, where the elements of $\mathfrak{m}^{\alpha_{ik}}$ are given degree $(1,i)$ for $1 \le k \le c(i)$.

The Lie algebra $\mathfrak{g}(A)$ satisfies the hypothesis of Theorem 2.1. Therefore, setting

$$\mathcal{V}_i = \coprod_{1\le k\le c(i)} U(\mathbb{C}f_{-1})\cdot f_{ik}, \tag{5}$$

$$\mathcal{V}'_i = \coprod_{1\le k\le c(i)} U(\mathbb{C}e_{-1})\cdot e_{ik} \tag{6}$$

for $i>0$ and defining

$$\mathcal{V} = \coprod_{i>0}\mathcal{V}_i \text{ and } \mathcal{V}' = \coprod_{i>0}\mathcal{V}'_i, \tag{7}$$

it follows (see [19]) that

$$\mathfrak{m} = \mathfrak{u}^+ \oplus \mathfrak{gl}_2 \oplus \mathfrak{u}^-, \tag{8}$$

with $\mathfrak{u}^- = L(\mathcal{V})$ and $\mathfrak{u}^+ = L(\mathcal{V}')$. In terms of the decomposition (4),

$$\mathfrak{u}^+ = \coprod_{m,n>0}\mathfrak{m}_{(m,n)},$$

$$\mathfrak{u}^- = \coprod_{m,n<0}\mathfrak{m}_{(m,n)}. \tag{9}$$

Note that the subspaces $\mathfrak{u}^\pm$ are $(\mathfrak{gl}_2, M)$-modules.

Lemma 2.3 *Let $\mathcal{V}$, $\mathcal{V}_i$ and $\mathfrak{u}^\pm$ be as above. The $\mathbb{Z}\oplus\mathbb{Z}$-graded vector spaces $\mathcal{V}$, $\mathcal{V}_i$ and $\mathfrak{u}^\pm$ are $(\mathfrak{gl}_2, M)$-modules. For $i>0$, let W_i denote the (unique up to isomorphism) irreducible $\mathfrak{gl}_2$-module of dimension i on which z acts as $i+1$. Then $\mathcal{V}_i \cong W_i \otimes V_i^\natural$, where W_i is regarded as a trivial M-module and $V_i^\natural$ as a trivial $\mathfrak{gl}_2$-module. Furthermore, as M-modules, $\mathcal{V}_i \cong iV_i^\natural$ and, for $m,n>0$,*

$$(\mathcal{V}_i)_{(-m,-n)} \cong \begin{cases} V_i^\natural & \text{if } m+n=i+1 \\ 0 & \text{otherwise.}\end{cases} \tag{10}$$

Proof: Clearly, the space of highest weight vectors of $\mathcal{V}_i$ has basis $\{f_{ik} \mid 1\le k\le c(i)\}$ and each f_{ik} has degree $(-1,-i)$. Since h acts on each f_{ik} as multiplication by $i-1$ and z acts as multiplication by $i+1$, each of these highest weight vectors generates a submodule isomorphic to W_i. Therefore,

$$\mathcal{V}_i \cong W_i \otimes \coprod_{1\le k\le c(i)}\mathbb{C}f_{ik}.$$

The isomorphism in [19] between $\mathfrak{g}(A)/\mathfrak{c}$ and $\mathfrak{m}$ is such that the subspace $\coprod_{1\le k\le c(i)}(\mathfrak{g}(A)/\mathfrak{c})^{\alpha_{ik}}$ is mapped isomorphically to $\mathfrak{m}_{(1,i)}$, so the images of the f_{ik} (respectively, the e_{ik}) for $1\le k\le c(i)$ form a basis for the space $\mathfrak{m}_{(-1,-i)}$

(respectively, $\mathfrak{m}_{(1,i)}$). Identifying the space $\mathfrak{m}_{(-1,-i)}$ with the M-module $V_i^\natural$ via Lemma 2.2, we have

$$\coprod_{1\le k\le c(i)} \mathbb{C}f_{ik} \cong V_i^\natural. \tag{11}$$

In this way we give the $\mathcal{V}_i$ (and therefore $\mathcal{V}$) M-module structures, so they are $(\mathfrak{gl}_2, M)$-modules. $\square$

Just as in the case of the copy of $\mathfrak{gl}_2 = \operatorname{span}\{e, f, h, z\}$ inside $\mathfrak{m} = P_1/R$ described above, it is easy to identify copies of the vector spaces $\mathcal{V}$ and $\mathcal{V}'$ inside $\mathfrak{m}$: For $i > 0$, $\mathfrak{m}_{(-1,-i)}$ consists of highest weight vectors for $\mathfrak{gl}_2$. The sum of all of the $\mathfrak{gl}_2$-modules generated by the $\mathfrak{m}_{(-1,-i)}$, $i > 0$, is a copy of $\mathcal{V}$ contained in $\mathfrak{m}$. The space $\mathcal{V}'$ can be analogously identified in $\mathfrak{m}$. We know that the Lie subalgebras of $\mathfrak{m}$ generated by $\mathcal{V}$ and $\mathcal{V}'$ are free Lie algebras. This suggests the following:

Conjecture: The associative algebra generated by the set $\{v_0 \mid v \in \mathcal{V}\}$ of weight-zero components of the vertex operators $Y(v, z)$ ($v \in \mathcal{V}$) and the associative algebra generated by $\{v_0 \mid v \in \mathcal{V}'\}$, acting on $\mathfrak{m}$, are free associative algebras. Equivalently, these algebras are isomorphic to $T(\mathcal{V})$ and $T(\mathcal{V}')$, respectively.

The Lie algebra $\mathfrak{m}$ is not the only quotient of $\mathfrak{g}(A)$ which has an action of M. We now construct another example. Let $\mathfrak{h}_i$, $i \in \mathbb{Z}_+$, be the span of the h_{ik} for $1 \le k \le c(i)$. Let $\mathfrak{c}_i = \mathfrak{c} \cap \mathfrak{h}_i$. Note that if $1 \le k, l \le c(i)$, then $h_{ik} - h_{il} \in \mathfrak{c}_i$ (since the corresponding columns of A are identical), and in fact these elements span $\mathfrak{c}_i$, so that $\dim \mathfrak{h}_i/\mathfrak{c}_i = 1$. Define $\mathfrak{c}' = \coprod_{i=1}^\infty \mathfrak{c}_i$, which is a central ideal of $\mathfrak{g}(A)$.

Proposition 2.4 *The action of M on the Lie algebra $\mathfrak{m}$ lifts naturally to an action of M on the Lie algebra $\mathfrak{g}(A)/\mathfrak{c}'$, in such a way that M acts trivially on $\mathfrak{h}/\mathfrak{c}'$.*

Proof: We identify the elements e_{ik} and f_{ik}, $(i, k) \in \mathcal{I}$, in the Lie algebras $\mathfrak{g}(A)$, $\mathfrak{g}(A)/\mathfrak{c}'$ and $\mathfrak{m}$. The invariant bilinear form on $\mathfrak{g}(A)$ is preserved, in the obvious sense, when we take quotients of $\mathfrak{g}(A)$ by central ideals. By (11), for each $i > 0$ we have that the span of the f_{ik} and the span of the e_{ik}, $1 \le k \le c(i)$, are M-modules. Recall that the e_{-1} and f_{-1} are fixed by the action of M. Since $h_{ik} - h_{il} \in \mathfrak{c}_i$ for $1 \le k, l \le c(i)$, we can write the image of each h_{ik} in $\mathfrak{g}(A)/\mathfrak{c}'$ as h_i. The Lie algebra $\mathfrak{g}(A)/\mathfrak{c}'$ then has a presentation with generators e_{ik}, f_{ik}, h_i, $i \in \mathcal{I}_0$, $1 \le k \le c(i)$ and relations induced by the relations (R1)-(R5) (the relation (R6) is vacuous) of $\mathfrak{g}(A)$:

(R1) $[h_i, h_j] = 0$,

(R2) $[h_l, e_{ik}] = -(l + i)e_{ik}$,

(R3) $[h_l, f_{ik}] = (l + i)f_{ik}$,

(R4) $[e_{ik}, f_{jm}] = (e_{ik}, f_{jm})h_i$,

(R5) $(\operatorname{ad} e_{-1})^i e_{ik} = 0$ and $(\operatorname{ad} f_{-1})^i f_{ik} = 0$

for $i, j \in \mathcal{I}_0$, $1 \leq k \leq c(i)$, $1 \leq m \leq c(j)$.

In $\mathfrak{g}(A)/\mathfrak{c}'$ let M act trivially on $\mathfrak{h}/\mathfrak{c}'$, so that M acts trivially on all the h_i. Relations (R1), (R2), (R3) and (R5) are M-invariant by inspection. The relation (R4) is equivalent to the assertion that for all $e \in \coprod_{1\leq k\leq c(i)} \mathbb{C}e_{ik}$ and $f \in \coprod_{1\leq k\leq c(i)} \mathbb{C}f_{ik}$,

$$[e, f] = (e, f)h_i.$$

The form $(\cdot, \cdot)$ is M-invariant on $\coprod_{1\leq k\leq c(i)} \mathbb{C}e_{ik} \times \coprod_{1\leq k\leq c(i)} \mathbb{C}f_{ik}$, because by Lemma 2.2, the form is M-invariant on $\mathfrak{m}$. Since $(e_{ik}, f_{lm}) = 0$ if $i \neq l$, we conclude that relation (R4) is also M-invariant. □

Of course, M also acts naturally on those generalized Kac-Moody algebras formed by taking further quotients of $\mathfrak{g}(A)/\mathfrak{c}'$ by central ideals.

2.4 Local Lie algebras

We now review some facts about local Lie algebras. These will be used (in Section 3) in the construction of certain generalized Kac-Moody algebras. The material in this subsection is applicable for an arbitrary field of characteristic zero.

Definition: A *local Lie algebra* is a graded vector space $\mathfrak{a} = \mathfrak{a}_{-1} \oplus \mathfrak{a}_0 \oplus \mathfrak{a}_1$ with skew-symmetric bilinear maps $[\cdot, \cdot]_{loc} : \mathfrak{a}_i \times \mathfrak{a}_j \to \mathfrak{a}_{i+j}$ for $|i|, |j|, |i + j| \leq 1$, satisfying:

i. $[\cdot, \cdot]_{loc} : \mathfrak{a}_0 \times \mathfrak{a}_0 \to \mathfrak{a}_0$ gives $\mathfrak{a}_0$ the structure of a Lie algebra.

ii. $[\cdot, \cdot]_{loc} : \mathfrak{a}_0 \times \mathfrak{a}_{\pm 1} \to \mathfrak{a}_{\pm 1}$ gives $\mathfrak{a}_{\pm 1}$ the structure of an $\mathfrak{a}_0$-module.

iii. $[\cdot, \cdot]_{loc} : \mathfrak{a}_{-1} \times \mathfrak{a}_1 \to \mathfrak{a}_0$ gives a map $\mathfrak{a}_{-1} \otimes \mathfrak{a}_1 \to \mathfrak{a}_0$ of $\mathfrak{a}_0$-modules.

Let $\mathfrak{l} = \coprod_{i\in\mathbb{Z}} \mathfrak{l}_i$ be a $\mathbb{Z}$-graded Lie algebra (with product $[\cdot, \cdot]$). Define $[\cdot, \cdot]_{loc} : \mathfrak{l}_i \times \mathfrak{l}_j \to \mathfrak{l}_{i+j}$ to be the restriction of $[\cdot, \cdot]$ to $\mathfrak{l}_i \times \mathfrak{l}_j$ whenever $|i|, |j|, |i + j| \leq 1$. Then $\mathfrak{l}_{-1} \oplus \mathfrak{l}_0 \oplus \mathfrak{l}_1$ has the structure of a local Lie algebra, which we call the *local part of* $\mathfrak{l}$ and denote by $\mathfrak{l}_{loc}$.

We say that two local Lie algebras $\mathfrak{a}, \mathfrak{b}$ are *isomorphic* if there is an isomorphism $\phi : \mathfrak{a} \to \mathfrak{b}$ of graded vector spaces such that $\phi[a_i, b_j]_{loc} = [\phi a_i, \phi b_j]_{loc}$ for all $a_i \in \mathfrak{a}_i, b_j \in \mathfrak{a}_j, |i|, |j|, |i + j| \leq 1$. A homomorphism of local Lie algebras is defined analogously.

Definition: A $\mathbb{Z}$-graded Lie algebra $\mathfrak{l} = \coprod_{i\in\mathbb{Z}} \mathfrak{l}_i$ is *associated to the local Lie algebra* $\mathfrak{a} = \mathfrak{a}_{-1} \oplus \mathfrak{a}_0 \oplus \mathfrak{a}_1$ if $\mathfrak{l}_{loc}$ is isomorphic to $\mathfrak{a}$, and $\mathfrak{l}$ is generated by $\mathfrak{l}_{-1} \oplus \mathfrak{l}_0 \oplus \mathfrak{l}_1$.

The following result is due to Kac [20, Propositions 4 and 7] in case $\mathfrak{a}$ is finite-dimensional over a field of characteristic zero. The same proof remains valid for arbitrary $\mathfrak{a}$.

Proposition 2.5 *Let $\mathfrak{a} = \mathfrak{a}_{-1} \oplus \mathfrak{a}_0 \oplus \mathfrak{a}_1$ be a local Lie algebra. Then up to isomorphism there is a unique $\mathbb{Z}$-graded Lie algebra $\mathfrak{a}_{max} = \coprod_{i\in\mathbb{Z}} \mathfrak{a}_i$ which is associated to $\mathfrak{a}$ and is maximal in the sense that any graded Lie algebra associated to $\mathfrak{a}$ is a quotient of $\mathfrak{a}_{max}$. Furthermore, $\mathfrak{a}^{\pm}_{max}$ is the free Lie algebra $L(\mathfrak{a}_{\pm 1})$ (recall the notation in Section 2.1). If $(\cdot,\cdot)_{loc}$ is a symmetric bilinear form on $\mathfrak{a}$ which is invariant with respect to $[\cdot,\cdot]_{loc}$, in the obvious sense, then $(\cdot,\cdot)_{loc}$ has a unique extension to an invariant symmetric bilinear form $(\cdot,\cdot)$ on $\mathfrak{a}_{max}$.*

3 Construction of a Lie algebra from a Kac-Moody algebra and a module

In this section we will give a construction of a Lie algebra starting from a Kac-Moody algebra and a suitable module for this algebra. We will show that every generalized Kac-Moody algebra in which no pair of distinct imaginary simple roots is orthogonal arises from our construction (so that in particular, the Monster Lie algebra does).

Let Γ be an index set, and let $V = \coprod_{\gamma\in\Gamma} V_\gamma$ be a Γ-graded vector space over $\mathbb{F}$, which for the general considerations in this section, through Corollary 3.1, we take to be an arbitrary field. For a subset $S \subset \Gamma$ let

$$V_{(S)} = \coprod_{\gamma\in\Gamma\setminus S} V_\gamma.$$

Clearly, $V_{(S)} \cap V_{(T)} = V_{(S\cup T)}$ and so $\{V_{(S)} | S \subset \Gamma, S\ finite\}$ forms a base of neighborhoods of 0 for a topology on V. As usual we give the base field the discrete topology. Let $V' \subset V^*$ denote the space of continuous linear functionals on V. Note that the topology of V and hence the space V' depend on the grading of V. Using a given involution of Γ, i.e., a map

$$\Gamma \longrightarrow \Gamma$$

$$\gamma \longmapsto \gamma'$$

of order one or two, we define a Γ-grading on V' by identifying $(V_\gamma)^*$ with

$$V'_{\gamma'} = \{v' \in V^* | v'(v) = 0\ for\ v \in V_{(\{\gamma\})}\}$$

for $\gamma \in \Gamma$; then

$$V' = \coprod_{\gamma\in\Gamma} V'_\gamma.$$

(Later, we shall take Γ to be an abelian group and γ' to be $-\gamma$.)

We say that a bilinear form $(\cdot,\cdot)$ on V is *compatible* with the grading and with a given involution $\gamma \mapsto \gamma'$ of Γ if $(V_\gamma, V_\mu) = 0$ if $\mu \neq \gamma'$. We say that V is *finite-dimensionally graded* if each V_γ is finite-dimensional.

Suppose that $(\cdot,\cdot)$ is a nondegenerate (symmetric) bilinear form on V which is compatible with the grading and with a given involution $\gamma \mapsto \gamma'$. Then the restriction of $(\cdot,\cdot)$ to each $V_\gamma \times V_{\gamma'}$ is nondegenerate. It follows that if V is finite-dimensionally graded, then for $f \in V'$ there exists a unique $v_f \in V$ such that

$$f(v) = (v_f, v)$$

for all $v \in V$.

Lemma 3.1 *Let $\mathfrak{a}_0$ be a Lie algebra finite-dimensionally graded by Γ (as a vector space) and let $\mathfrak{a} = \mathfrak{a}_{-1} \oplus \mathfrak{a}_0 \oplus \mathfrak{a}_1$ be an $\mathfrak{a}_0$-module graded by $\{-1, 0, 1\}$. Let $(\cdot,\cdot)$ be an $\mathfrak{a}_0$-invariant symmetric bilinear form on $\mathfrak{a}$. Assume that $(\mathfrak{a}_i, \mathfrak{a}_j) = 0$ whenever $i + j \neq 0$, that the restriction of $(\cdot,\cdot)$ to $\mathfrak{a}_0 \times \mathfrak{a}_0$ is nondegenerate and compatible with the grading on $\mathfrak{a}_0$ and with a given involution on Γ, and that for any $a_{\pm 1} \in \mathfrak{a}_{\pm 1}$ the linear functional*

$$a_0 \longmapsto (a_{-1}, a_0 \cdot a_1)$$

is continuous on $\mathfrak{a}_0$. Then there is a unique local Lie algebra structure on $\mathfrak{a}$ which extends the $\mathfrak{a}_0$-module structure of $\mathfrak{a}$ and such that $(\cdot,\cdot)$ is invariant with respect to this structure.

Proof: Let $a_i \in \mathfrak{a}_i$ for $i = -1, 0, 1$. We (must) define

$$[a_0, a_{\pm 1}]_{loc} = -[a_{\pm 1}, a_0]_{loc} = a_0 \cdot a_{\pm 1}.$$

The rest of $[\cdot,\cdot]_{loc}$ is determined uniquely as follows: By hypothesis, the linear functional

$$\begin{aligned} &\mathfrak{a}_0 \longrightarrow \mathbb{F} \\ &a_0 \longmapsto (a_{-1}, [a_1, a_0]_{loc}) \end{aligned}$$

is continuous. Since $(\cdot,\cdot)$ is nondegenerate and compatible with the grading on $\mathfrak{a}_0$ there is a unique element of $\mathfrak{a}_0$, which we denote by $[a_{-1}, a_1]_{loc}$, such that

$$([a_{-1}, a_1]_{loc}, a_0) = (a_{-1}, [a_1, a_0]_{loc}) \tag{12}$$

for all $a_0 \in \mathfrak{a}_0$. Clearly, $[\cdot,\cdot]_{loc}$ is bilinear on $\mathfrak{a}_{-1} \times \mathfrak{a}_1$. Set $[a_1, a_{-1}]_{loc} = -[a_{-1}, a_1]_{loc}$.

Then for any $a_0, a_0' \in \mathfrak{a}_0$ we have

$$\begin{aligned} ([[a_{-1}, a_1]_{loc}, a_0]_{loc}, a_0') &= ([a_{-1}, a_1]_{loc}, [a_0, a_0']_{loc}) = (a_{-1}, [a_1, [a_0, a_0']_{loc}]_{loc}) \\ &= (a_{-1}, [[a_1, a_0]_{loc}, a_0']_{loc}) + (a_{-1}, [a_0, [a_1, a_0']_{loc}]_{loc}) \\ &= ([a_{-1}, [a_1, a_0]_{loc}]_{loc}, a_0') + ([a_{-1}, a_0]_{loc}, [a_1, a_0']_{loc}) \\ &= ([a_{-1}, [a_1, a_0]_{loc}]_{loc}, a_0') + ([[a_{-1}, a_0]_{loc}, a_1]_{loc}, a_0'). \end{aligned}$$

In view of the nondegeneracy of $(\cdot,\cdot)|_{\mathfrak{a}_0\times\mathfrak{a}_0}$ this implies

$$[[a_{-1},a_1]_{loc},a_0]_{loc}=[a_{-1},[a_1,a_0]_{loc}]_{loc}+[[a_{-1},a_0]_{loc},a_1]_{loc}.$$

This completes the proof that $[\cdot,\cdot]_{loc}$ gives $\mathfrak{a}$ the structure of a local Lie algebra. The invariance of $(\cdot,\cdot)$ now follows from (12) together with the $\mathfrak{a}_0$-invariance of $(\cdot,\cdot)$, the skew symmetry of $[\cdot,\cdot]_{loc}$ and the symmetry of $(\cdot,\cdot)$. $\square$

Now let $W=\coprod W_i$ be a direct sum of finite-dimensionally graded vector spaces. For each i, identify W_i^* with $\{w^*\in W^*|w^*(w)=0\ for\ w\in W_j, j\neq i\}$. Then $W_i'\subset W_i^*$ (defined above) is identified with a subspace of W^*. We set

$$W'=\coprod W_i'\subset W^*.$$

For $w\in W, w'\in W'$ define $(w',w)=(w,w')=w'(w)$. Also set $(W,W)=(W',W')=0$. Thus $(\cdot,\cdot)$ is a nondegenerate symmetric bilinear form on $W\oplus W'$.

Corollary 3.2 *Let $\mathfrak{a}_0$ be a Lie algebra which is finite-dimensionally graded, as a Lie algebra, by an abelian group Γ. Assume that $\mathfrak{a}_0$ has a nondegenerate invariant symmetric bilinear form $(\cdot,\cdot)$ compatible with the grading and the involution $\gamma\mapsto-\gamma$ of the abelian group Γ. Let $W=\coprod W_i$ be a direct sum of finite-dimensionally Γ-graded $\mathfrak{a}_0$-modules. Then there is a unique $\mathfrak{a}_0$-module structure on W' such that the nondegenerate symmetric bilinear form $(\cdot,\cdot)$ on $W\oplus W'$ defined above is $\mathfrak{a}_0$-invariant, and W' is Γ-graded as an $\mathfrak{a}_0$-module. Set $\mathfrak{a}_1=W,\mathfrak{a}_{-1}=W',\mathfrak{a}=\mathfrak{a}_{-1}\oplus\mathfrak{a}_0\oplus\mathfrak{a}_1$. Let $(\cdot,\cdot)$ be the nondegenerate $\mathfrak{a}_0$-invariant symmetric bilinear form on $\mathfrak{a}$ which extends the forms $(\cdot,\cdot)$ on $\mathfrak{a}_0$ and on $W\oplus W'$, so that $(\mathfrak{a}_i,\mathfrak{a}_j)=0$ if $i+j\neq0$, and $(a_{-1},a_1)=a_{-1}(a_1)$ for $a_{\pm1}\in\mathfrak{a}_{\pm1}$. Then $\mathfrak{a}$ has a unique local Lie algebra structure extending the $\mathfrak{a}_0$-module structure of $\mathfrak{a}$ such that $(\cdot,\cdot)$ is invariant.*

Proof: To prove the first assertion, we first note that for each i, the bilinear form $(\cdot,\cdot)$ on $W_i\oplus W_i'$ is compatible with the grading and the involution $\gamma\mapsto-\gamma$. For $a_0\in\mathfrak{a}_0$ and $w_i'\in W_i'$, the linear functional

$$\varphi: w_i\mapsto(w_i',a_0\cdot w_i)$$

is continuous. In fact, it is sufficient to show this for a_0 in some $(\mathfrak{a}_0)_\mu$ and w_i' in some $(W_i')_\nu$, in which case φ vanishes on $(W_i)_\gamma$ unless $\gamma+\mu+\nu=0$. Thus φ is continuous. We (must) define $a_0\cdot w_i'\in W_i'$ so that

$$(a_0\cdot w_i',w_i)=-(w_i',a_0\cdot w_i),$$

and $a_0\cdot w_i'$ is clearly bilinear in a_0 and w_i', and clearly defines an $\mathfrak{a}_0$-module action on W'. It is straightforward to verify that W' is Γ-graded as an $\mathfrak{a}_0$-module. This completes the proof of the first assertion.

To prove the remaining assertion, we use Lemma 3.1. To apply this, we must show that the linear functional

$$\psi : a_0 \longmapsto (a_{-1}, a_0 \cdot a_1)$$

is continuous for all $a_{\pm 1} \in \mathfrak{a}_{\pm 1}$. It is sufficient to verify this for a_1 in some $(W_i)_\mu$ and a_{-1} in some $(W_j')_\nu$. Then ψ vanishes on $\mathfrak{a}_\gamma$ unless $\gamma + \mu + \nu = 0$ and hence ψ is continuous, as required. $\square$

Now we return to the setting of Section 2.2. Let S be a finite set and let $B = (b_{ij})_{i,j \in S}$ be a matrix satisfying the conditions (C1)-(C3) and the additional condition $b_{ii} > 0$ for all $i \in S$. Then $\mathfrak{g}(B)^e$ is the (symmetrizable) Kac-Moody algebra associated to B. Recall that the symmetric invariant bilinear form $(\cdot, \cdot)_{\mathfrak{g}(B)^e}$ on $\mathfrak{g}(B)^e$ is nondegenerate and satisfies the condition $(e_i, f_j)_{\mathfrak{g}(B)^e} = \delta_{ij}$ for all $i, j \in S$.

If V is an (irreducible) integrable lowest weight $\mathfrak{g}(B)^e$-module (necessarily finite-dimensionally graded by its weight-space decomposition) with lowest weight λ, then V' carries the structure of an integrable highest weight module with highest weight $-\lambda$.

Let $\mathfrak{z}$ be a finite-dimensional abelian Lie algebra and let $\mathfrak{a}_0 = \mathfrak{g}(B)^e \oplus \mathfrak{z}$, a direct sum of Lie algebras.

Let $W = \coprod_{j \in J} W_j$ be a direct sum of $\mathfrak{a}_0$-modules, where each W_j is an integrable lowest weight $\mathfrak{g}(B)^e$-module, with lowest weight μ_j and lowest weight vector e_j. (We assume that the two index sets S and J are disjoint.) Since each element of $\mathfrak{z}$ must act as a scalar on W_j, there is some $\gamma_j \in \mathfrak{z}^*$ such that $z \cdot w = \gamma_j(z) w$ for all $z \in \mathfrak{z},\ w \in W_j$. Assume for convenience that

$$\bigcap_{j \in J} \operatorname{Ker} \gamma_j = 0.$$

Let $(\cdot, \cdot)_\mathfrak{z}$ be a nonsingular symmetric bilinear form on $\mathfrak{z}$ (which is necessarily $\mathfrak{z}$-invariant). Define $(\cdot, \cdot)$ on $\mathfrak{a}_0$ by $(g + z, g' + z') = (g, g')_{\mathfrak{g}(B)^e} - (z, z')_\mathfrak{z}$ for $g, g' \in \mathfrak{g}(B)^e, z, z' \in \mathfrak{z}$. Note that this form is symmetric and nondegenerate.

Now $W' = \coprod_{j \in J} W_j'$ (see above) and each W_j' is an irreducible integrable highest weight $\mathfrak{g}(B)^e$-module, with highest weight $-\mu_j$ and a highest weight vector f_j which we choose so as to satisfy the condition $(e_j, f_j) = 1$. Clearly, $z \cdot w' = -\gamma_j(z) w'$ for all $z \in \mathfrak{z}$ and $w' \in W_j'$. Also, $(e_i, f_j) = \delta_{ij}$ for $i, j \in J$.

Set $\mathfrak{k} = \mathfrak{h}^e \oplus \mathfrak{z}$ and note that $(\cdot, \cdot)$ is nonsingular on $\mathfrak{k}$. We identify $\mathfrak{k}^*$ with $(\mathfrak{h}^e)^* \oplus \mathfrak{z}^*$. Let $(\cdot, \cdot)_{\mathfrak{h}^e}$ denote the restriction of the form $(\cdot, \cdot)_{\mathfrak{g}(B)^e}$ to $\mathfrak{h}^e$. Then for $\mu \in (\mathfrak{h}^e)^*$ there exists a unique $h_\mu \in \mathfrak{h}^e$ such that $(h_\mu, h)_{\mathfrak{h}^e} = \mu(h)$ for all $h \in \mathfrak{h}^e$ and for $\gamma \in \mathfrak{z}^*$ there exists a unique $z_\gamma \in \mathfrak{z}$ such that $(z_\gamma, z)_\mathfrak{z} = \gamma(z)$ for all $z \in \mathfrak{z}$. We denote $h_\mu + z_\gamma$ by $k_{\mu - \gamma} \in \mathfrak{k}$ and note that $(k_{\mu-\gamma}, k) = (\mu - \gamma)(k)$ for all $k \in \mathfrak{k}$. As usual, we may transfer the nonsingular forms $(\cdot, \cdot)_{\mathfrak{h}^e}$ to $(\mathfrak{h}^e)^*$ and $(\cdot, \cdot)_\mathfrak{z}$ to $\mathfrak{z}^*$ by defining $(\alpha, \beta)_{(\mathfrak{h}^e)^*} = (h_\alpha, h_\beta)_{\mathfrak{h}^e}$ for $\alpha, \beta \in (\mathfrak{h}^e)^*$ and $(\gamma, \sigma)_{\mathfrak{z}^*} = (z_\gamma, z_\sigma)_\mathfrak{z}$ for $\gamma, \sigma \in \mathfrak{z}^*$. There is a corresponding form $(\cdot, \cdot)$ on $\mathfrak{k}^*$ defined by $(\cdot, \cdot)_{(\mathfrak{h}^e)^*} - (\cdot, \cdot)_{\mathfrak{z}^*}$.

Thus $(\nu, \tau) = (k_\nu, k_\tau)$ for $\nu, \tau \in \mathfrak{k}^*$. Note that

$$\mathfrak{z} = \text{span}\,\{z_{\gamma_j} \mid j \in J\}.$$

With the $\mathfrak{a}_0$ given above and $\Gamma = \mathfrak{k}^*$, let

$$\mathfrak{a} = W' \oplus \mathfrak{a}_0 \oplus W = \mathfrak{a}_{-1} \oplus \mathfrak{a}_0 \oplus \mathfrak{a}_1$$

be the local Lie algebra given by Corollary 3.2 and let $\mathfrak{a}_{max}$ be the associated $\mathbb{Z}$-graded Lie algebra given by Proposition 2.5.

Next we give some basic bracket relations together with relevant definitions:

Proposition 3.3 *The following relations hold in* $\mathfrak{a}$*:*

(a) If $i, j \in S$ *then* $[e_i, f_j] = \delta_{ij} h_i = \delta_{ij} h_{\alpha_i}$.

(b) If $i, j \in J$ *then* $[e_i, f_j] = \delta_{ij} k_i$*, where*

$$k_i = k_{\mu_i + \gamma_i}.$$

(c) If $i \in S$ *and* $j \in J$ *then* $[e_i, f_j] = 0$ *and* $[f_i, e_j] = 0$.

Proof: (a) is clear.

(b) For $i, j \in J$, $[e_i, f_j]$ is an element of $\mathfrak{a}_0$ which is orthogonal to the subalgebras $\mathfrak{n}^{\pm} \subset \mathfrak{g}(B)^e$, so that $[e_i, f_j] \in \mathfrak{k}$. Now let $k \in \mathfrak{k}$. Then $(k, [e_i, f_j]) = ([k, e_i], f_j) = ((\mu_i(k) + \gamma_i(k))e_i, f_j) = \delta_{ij}(\mu_i(k) + \gamma_i(k))$, so $[e_i, f_j] = \delta_{ij} k_{\mu_j + \gamma_j}$.

(c) follows from the definitions of e_j and f_j. $\square$

Proposition 3.4 *The following relations hold in* $\mathfrak{a}$*:*

(a) Let $i, j \in S$*. Then* $[h_i, e_j] = a_{ij} e_j$ *and* $[h_i, f_j] = -a_{ij} f_j$*, where*

$$a_{ij} = b_{ij} = (\alpha_i, \alpha_j).$$

(b) Let $i \in S, j \in J$*. Then* $[h_i, e_j] = a_{ij} e_j$*,* $[h_i, f_j] = -a_{ij} f_j$*,* $[k_j, e_i] = a_{ji} e_i$*, and* $[k_j, f_i] = -a_{ji} f_j$*, where*

$$a_{ij} = a_{ji} = (\alpha_i, \mu_j).$$

(c) Let $i, j \in J$*. Then* $[k_i, e_j] = a_{ij} e_j$ *and* $[k_i, f_j] = -a_{ij} f_j$*, where*

$$a_{ij} = (\mu_i, \mu_j)_{(\mathfrak{h}^e)^*} - (\gamma_i, \gamma_j)_{\mathfrak{z}^*}.$$

Proof: (a) is clear.

(b) We have

$$[h_i, e_j] = \mu_j(h_i) e_j = \mu_j(h_{\alpha_i}) e_j = (\alpha_i, \mu_j) e_j$$

and

$$[k_j, e_i] = [k_{\mu_j+\gamma_j}, e_i] = \alpha_i(h_{\mu_j})e_i = (\mu_j, \alpha_i)e_i.$$

The other cases are similar.

(c) We have

$$\begin{aligned}[k_i, e_j] &= [k_{\mu_i+\gamma_i}, e_j] = [h_{\mu_i} - z_{\gamma_i}, e_j] \\ &= (\mu_j(h_{\mu_i}) - \gamma_j(z_{\gamma_i}))e_j = ((\mu_i, \mu_j)_{(\mathfrak{h}^e)^*} - (\gamma_i, \gamma_j)_{\mathfrak{z}^*})e_j,\end{aligned}$$

and the remaining case is similar. □

Let $A = (a_{ij})_{i,j\in S\cup J}$, where the a_{ij} are defined in the above proposition. Let $\mathfrak{d}_B$ be the derivation space associated with $\mathfrak{g}(B)$, i.e., $\mathfrak{g}(B)^e = \mathfrak{d}_B \ltimes \mathfrak{g}(B)$. Note that the containment $\mathfrak{d} \ltimes \mathfrak{g}(A) \subset \mathfrak{g}(A)^e$ can be strict. We assume that the restriction $(\cdot,\cdot)$ to $\mathfrak{g}(B)^e$ of the symmetric bilinear form $(\cdot,\cdot)$ on $\mathfrak{g}(A)^e$ has all the usual properties, in particular, that it is nondegenerate. The radical $\mathrm{Rad}(\cdot,\cdot)$ of the form $(\cdot,\cdot)$ on $\mathfrak{d}_B \ltimes \mathfrak{g}(A)$ is central in $\mathfrak{g}(A)^e$.

Theorem 3.5

(a) Assume that $a_{ij} = (\mu_i, \mu_j)_{(\mathfrak{h}^e)^*} - (\gamma_i, \gamma_j)_{\mathfrak{z}^*} \leq 0$ *for all* $i, j \in J$. *Then* A *satisfies* (C1)-(C3) *and the Lie algebra* $(\mathfrak{d}_B \ltimes \mathfrak{g}(A))/\mathrm{Rad}(\cdot,\cdot)$ *is a Lie algebra associated to the local Lie algebra* $\mathfrak{a}$. *In particular, this Lie algebra is a quotient of* $\mathfrak{a}_{max}$.

(b) Assume that $(\mu_i, \mu_j)_{(\mathfrak{h}^e)^*} - (\gamma_i, \gamma_j)_{\mathfrak{z}^*} \leq 0$ *for all* $i, j \in J$ *and that* $(\mu_i, \mu_j)_{(\mathfrak{h}^e)^*} - (\gamma_i, \gamma_j)_{\mathfrak{z}^*} < 0$ *for all* $i \neq j$. *Then* $(\mathfrak{d}_B \ltimes \mathfrak{g}(A))/\mathrm{Rad}(\cdot,\cdot) \cong \mathfrak{a}_{max}$.

Proof: We first prove (a). Since W_j is an integrable lowest weight module with lowest weight μ_j, the properties (C1)-(C3) are clear.

We shall use the symbols e_i^A, f_i^A and h_i^A, $i \in S \cup J$, to denote the generators of $\mathfrak{g}(A)$, and $\mathfrak{h}^A$ will denote the Cartan subalgebra of $\mathfrak{g}(A)$. Using the notation defined in (3) above, we have $\mathfrak{r} = \mathfrak{g}(B) + \mathfrak{h}^A$ and $\mathfrak{g}(A) = \mathfrak{u}^+ \oplus \mathfrak{r} \oplus \mathfrak{u}^-$. Thus $\mathfrak{d}_B \ltimes \mathfrak{g}(A) = \mathfrak{u}^+ \oplus \mathfrak{r}^e \oplus \mathfrak{u}^-$, where $\mathfrak{r}^e = \mathfrak{d}_B \ltimes \mathfrak{r}$. We give the Lie algebra $\mathfrak{d}_B \ltimes \mathfrak{g}(A)$ the $\mathbb{Z}$-grading determined by:

$$\deg e_j = -\deg f_j = \begin{cases} 0 & \text{if } j \in S \\ 1 & \text{if } j \in J. \end{cases}$$

Note that $\mathrm{Rad}(\cdot,\cdot) \subset (\mathfrak{d}_B \ltimes \mathfrak{g}(A))_0$, so that we can identify the elements e_i^A and f_i^A with their images in $(\mathfrak{d}_B \ltimes \mathfrak{g}(A))/\mathrm{Rad}(\cdot,\cdot)$. Let $\mathfrak{l} = (\mathfrak{d}_B \ltimes \mathfrak{g}(A))/\mathrm{Rad}(\cdot,\cdot)$. As $\mathfrak{d}_B \cap \mathrm{Rad}(\cdot,\cdot) = 0$, we can also identify $\mathfrak{d}_B$ with its image in $\mathfrak{l}$. The $\mathbb{Z}$-grading defined on $\mathfrak{d}_B \ltimes \mathfrak{g}(A)$ induces a $\mathbb{Z}$-grading on $\mathfrak{l}$.

Then $\mathfrak{l}_1 = \coprod_{j\in J} U(\mathfrak{n}_S^+)e_j^A$, $\mathfrak{l}_{-1} = \coprod_{j\in J} U(\mathfrak{n}_S^-)f_j^A$ and $\mathfrak{l}_0 = \mathfrak{r}^e/\mathrm{Rad}(\cdot,\cdot)$. The subspace $\mathfrak{l}_{loc} = \mathfrak{l}_{-1} \oplus \mathfrak{l}_0 \oplus \mathfrak{l}_1$ is the local part of the Lie algebra $\mathfrak{l}$ ($\mathfrak{l}_{loc}$ generates $\mathfrak{l}$ since it contains $\mathfrak{d}_B$ and the generators e_i^A and f_i^A, $i \in S \cup J$).

Thus, by Proposition 2.5, to prove that $\mathfrak{l}$ is a quotient of $\mathfrak{a}_{max}$ it is sufficient to prove $\mathfrak{l}_{loc} \cong \mathfrak{a}$. By the uniqueness assertion in Lemma 3.1, it is enough to give a linear isomorphism

$$\varphi : \mathfrak{l}_{loc} \to \mathfrak{a}$$

preserving the gradings by the set $\{-1, 0, 1\}$, such that for $x \in \mathfrak{l}_0$, $y \in \mathfrak{l}_{loc}$,

$$[\varphi(x), \varphi(y)] = \varphi([x, y]) \tag{13}$$

and such that for all $x, y \in \mathfrak{l}_{loc}$,

$$(\varphi(x), \varphi(y)) = (x, y). \tag{14}$$

To do this we will define a surjective homomorphism of graded vector spaces, which we still call φ,

$$\varphi : \mathfrak{l}_{-1} \oplus \mathfrak{r}^e \oplus \mathfrak{l}_1 \to \mathfrak{a}_{-1} \oplus \mathfrak{a}_0 \oplus \mathfrak{a}_1,$$

and show that (13) and (14) are satisfied and that the kernel of φ is $\mathrm{Rad}(\cdot, \cdot)$.

We define φ as a linear map by prescribing φ on the subspaces $\mathfrak{l}_{\pm 1}$, $\mathfrak{g}(B)^e$ and the span of the h_j^A for $j \in J$.

Each W_j for $j \in J$ is an integrable lowest weight $\mathfrak{g}(B)^e$-module of weight μ_j and so is isomorphic to the integrable lowest weight $\mathfrak{g}(B)^e$-module $U(\mathfrak{n}_S^+)e_j^A$ of weight μ_j by an isomorphism φ which takes the lowest weight vector e_j^A to e_j. This gives us $\varphi : \mathfrak{l}_1 \xrightarrow{\sim} \mathfrak{a}_1$. The isomorphism $\varphi : \mathfrak{l}_{-1} \xrightarrow{\sim} \mathfrak{a}_{-1}$ is defined analogously.

We define φ to be the identity on $\mathfrak{g}(B)^e$.

To define φ on all of $\mathfrak{r}^e$, all that remains is to define φ on $h_j^A \in \mathfrak{h}^A$ with $j \in J$. Set $\varphi(h_j^A) = k_j = k_{\mu_j + \gamma_j} \in \mathfrak{k}$.

This defines a homomorphism of graded vector spaces $\varphi : \mathfrak{l}_{-1} \oplus \mathfrak{r}^e \oplus \mathfrak{l}_1 \to \mathfrak{a}$. That φ satisfies (13) follows from Proposition 3.4. This map is surjective because the z_{γ_j} span $\mathfrak{z}$.

It is clear from the definition of the forms that $(x, y) = (\varphi(x), \varphi(y))$ for $x, y \in \mathfrak{g}(B)^e$. The invariant bilinear forms on $\mathfrak{l}_{-1} \oplus \mathfrak{r}^e \oplus \mathfrak{l}_1$ and on $\mathfrak{a}$ satisfy the relations

$$(e_i^A, f_j^A) = (e_i, f_j) = \delta_{ij}$$

for all $i, j \in J$. Since an invariant pairing between a lowest and highest weight $\mathfrak{g}(B)^e$-module is determined by its value on the pair consisting of the lowest and highest weight vectors, we conclude that $(x, y) = (\varphi(x), \varphi(y))$ for $x \in \mathfrak{l}_1$ and $y \in \mathfrak{l}_{-1}$. Furthermore, for $i \in S \cup J, j \in J$

$$\begin{aligned}
(h_i^A, h_j^A) = a_{ij} &= \begin{cases} (\alpha_i, \mu_j + \gamma_j) & \text{if } i \in S \\ (\mu_i + \gamma_i, \mu_j + \gamma_j) & \text{if } i \in J \end{cases} \\
&= \begin{cases} (h_{\alpha_i}, k_j) & \text{if } i \in S \\ (k_i, k_j) & \text{if } i \in J \end{cases} \\
&= (\varphi(h_i), \varphi(h_j)).
\end{aligned}$$

Also, it is clear that for $d \in \mathfrak{d}_B$ and $j \in J$, $(d, h_j^A) = (\varphi(d), \varphi(h_j^A))$. It follows that for all $x, y \in \mathfrak{l}_{-1} \oplus \mathfrak{r}^e \oplus \mathfrak{l}_1$, we have $(x, y) = (\varphi(x), \varphi(y))$.

The kernel of φ clearly lies in $\mathfrak{d}_B \oplus \mathfrak{h}^A$. Let h be an element of the kernel. Then for all $x \in \mathfrak{l}_{-1} \oplus \mathfrak{r}^e \oplus \mathfrak{l}_1$, $(h, x) = (\varphi(h), \varphi(x)) = 0$, so that $h \in \mathrm{Rad}(\cdot, \cdot)$. Conversely, if $y \in \mathrm{Rad}(\cdot, \cdot)$, then $(\varphi(y), a) = 0$ for all $a \in \mathfrak{a}$, since φ is surjective. As the bilinear form is nondegenerate on $\mathfrak{a}_0$, this means that $\varphi(y) = 0$. We conclude that the kernel of the map φ is $\mathrm{Rad}(\cdot, \cdot)$, completing the proof of (a).

Part (b) then follows by Theorem 2.1 and the definition of $\mathfrak{d}_B \ltimes \mathfrak{g}(A)$. $\square$

Remark 3.6 After possible reordering of rows and columns, every finite rank matrix satisfying (C1)-(C3) and having only finitely many positive diagonal entries arises as the matrix A for a suitable choice of the generalized Cartan matrix B, the module W and the elements γ_i. In fact, let $P = (p_{kl})_{k,l \in K}$ be any matrix as indicated. Let $S = \{i \in K | p_{ii} > 0\}$ and $J = \{j \in K | p_{jj} \leq 0\}$, so that S is a finite set. Reorder the rows and columns of P so that the rows and columns indexed by elements of S precede those indexed by elements of J. Let $B = (p_{kl})_{k,l \in S}$. Consider the Kac-Moody algebra $\mathfrak{g}(B)^e$. For $j \in J$, define μ_j to be any element of $(\mathfrak{h}^e)^*$ such that $\mu_j(h_i) = p_{ij}$ for $i \in S$. Let W_j be the integrable lowest weight $\mathfrak{g}(B)^e$-module with lowest weight μ_j. Define $q_{ij} = (\mu_i, \mu_j) - p_{ij}$ for $i, j \in J$ and let $Q = (q_{ij})_{i,j \in J}$. Let $\mathfrak{z}$ be an abelian Lie algebra of (finite) dimension equal to the rank of Q, equipped with a nonsingular symmetric bilinear form $(\cdot, \cdot)_{\mathfrak{z}}$. Let $(\cdot, \cdot)_{\mathfrak{z}^*}$ be the corresponding form on $\mathfrak{z}^*$. Let $\mathfrak{t}$ be a vector space over $\mathbb{C}$ with basis $\{\eta_j \mid j \in J\}$. Define an inner product on $\mathfrak{t}$ by $(\eta_i, \eta_j) = q_{ij}$ for all $i, j \in J$. Then $\mathfrak{t}/\mathrm{Rad}(\cdot, \cdot)$ has dimension equal to rank Q, hence to $\dim \mathfrak{z}^*$. We may therefore identify $\mathfrak{t}/\mathrm{Rad}(\cdot, \cdot)$ with $\mathfrak{z}^*$. Let γ_j denote the image of η_j under this identification, so that $(\gamma_i, \gamma_j)_{\mathfrak{z}^*} = q_{ij}$ for all $i, j \in J$. For these choices of B, $W = \coprod_{j \in J} W_j$ and the γ_j, the resulting matrix A is equal to P.

We give such a construction of the Monster Lie algebra explicitly: In this case (using the notation of Section 2.3), $K = \mathcal{I}$, $S = \{(-1, 1)\}$, $J = \{(i, k) | i \in \mathbb{Z}_+, 1 \leq k \leq c(i)\}$ and $B = (2)$. Thus $\mathfrak{g}(B) = \mathfrak{g}(B)^e = \mathfrak{sl}_2$, so that $\mathfrak{d}_B = 0$ and $\dim \mathfrak{h}^e = 1$. As before, we write h_{-1} for $h_{-1,1}$. Since for $i \in \mathbb{Z}_+$ $\mu_{(i,k)}$ depends only on i, we may denote it μ_i. Then $\mu_i(h_{-1}) = -(i-1)$ and $(\mu_i, \mu_j) = (i-1)(j-1)/2$. Hence the entries in the (i, j)-block of Q have value $(i+1)(j+1)/2$. We may therefore take $\mathfrak{z}$ to be one-dimensional. If z spans $\mathfrak{z}$ and satisfies $(z, z)_{\mathfrak{z}} = 1$, then we have $\gamma_i(z) = (i+1)/\sqrt{2}$. Note that $\mathfrak{a}_0 = \mathfrak{sl}_2 \oplus \mathfrak{z} \cong \mathfrak{gl}_2$. Because the center $\mathfrak{c}$ is equal to $\mathrm{Rad}(\cdot, \cdot)$, we conclude that $\mathfrak{m} = \mathfrak{a}_{max}$ for $\mathfrak{a} = \mathcal{V}' \oplus \mathfrak{gl}_2 \oplus \mathcal{V}$.

4 Generalized Verma modules and standard modules

Given a generalized Kac-Moody Lie algebra, we observe that certain standard modules are equal to induced modules. For Lie algebras $\mathfrak{g}$ satisfying the condi-

tions of Theorem 2.1, this means that the tensor algebra over a certain vector space has the structure of an irreducible module for $\mathfrak{g}$. In this way we show that $\mathfrak{m}$ (and $\mathfrak{g}(A)/\mathfrak{c}'$) can be realized naturally as an explicitly prescribed M-covariant Lie algebra of operators on $T(\mathcal{V})$ (which has the form of an induced $\mathfrak{m}$- or $\mathfrak{g}(A)/\mathfrak{c}'$-module for certain weights λ) for our M-module $\mathcal{V}$.

4.1 Examples of standard modules

The conditions in the definition below of standard module appear in [2]; the definition was made in [18]. Standard modules are the only modules for which a character formula is proven ([2]; see [17], [18] and [21]). Note that standard modules are not the same as "integrable highest weight modules" for generalized Kac-Moody algebras (see [21] for the definition of "integrable" in the context of generalized Kac-Moody algebras).

Let X be a $\mathfrak{g}^e$-module. A weight $\lambda \in (\mathfrak{h}^e)^*$ is *dominant* if $(\lambda, \alpha_i) \in \mathbb{R}$ and $(\lambda, \alpha_i) \geq 0$ for all $i \in I$. A weight λ is called *integral* if $\frac{2\lambda(h_i)}{a_{ii}} \in \mathbb{Z}$ for all $i \in I$ such that $a_{ii} > 0$. Denote by P_+ the set of dominant integral weights.

As in [18], define X to be a *standard* module if X is a highest weight module ([2] and [18] actually use lowest weight modules) with highest weight $\mu \in P_+$ and highest weight vector x such that:

1. for $i \in I$, if $(\mu, \alpha_i) = 0$ then $f_i \cdot x = 0$;
2. if α_i $(i \in I)$ is real then $f_i^{n_i+1} \cdot x = 0$, where $n_i = 2(\mu, \alpha_i)/(\alpha_i, \alpha_i)$ (necessarily a nonnegative integer).

Proposition 4.1 *All standard modules are irreducible.*

Proof: This follows from the character formula for a standard module stated in [2] and [21]. For two different complete proofs of the character formula see [17] and [18]. □

Let $\lambda \in P_+$, and consider the standard (irreducible) highest weight $\mathfrak{r}^e$-module $L(\lambda)$ associated to λ; it is irreducible as a $\mathfrak{g}_S^e$-module (cf. [21]). The highest weight space of $L(\lambda)$ (as a $\mathfrak{g}_S^e$-module) is a weight space for $\mathfrak{h}^e$, with weight λ. Let $\mathfrak{u}^+$ act trivially on $L(\lambda)$; this gives $L(\lambda)$ the structure of an irreducible $\mathfrak{p}^e$-module. Define the *generalized Verma module* $V^{L(\lambda)}$ (cf. [25]) to be the induced module $U(\mathfrak{g}^e) \otimes_{U(\mathfrak{p}^e)} L(\lambda)$. It is clear that $V^{L(\lambda)} \cong U(\mathfrak{u}^-) \otimes L(\lambda)$ as vector spaces.

We make the following observation:

Proposition 4.2 *Let $\lambda \in P_+$ be an element satisfying $(\lambda, \alpha_i) > 0$ for every imaginary simple root α_i $(i \in I \backslash S)$. Then the generalized Verma module $V^{L(\lambda)}$ is a standard, and therefore irreducible, module for $\mathfrak{g}(A)^e$.*

Proof: The module $V^{L(\lambda)}$ is a highest weight module with highest weight λ, and clearly satisfies conditions 1 and 2 above. $\square$

Proposition 4.2 immediately gives a class of examples of standard modules for arbitrary generalized Kac-Moody algebras. In the case of generalized Kac-Moody algebras satisfying the hypothesis of Theorem 2.1, like $\mathfrak{m}$, these generalized Verma modules are particularly easy to describe. As another example, we show that the irreducible module constructed in [15] for a generalized Kac-Moody Lie algebra with one imaginary simple root is an example of such a generalized Verma module.

Theorem 4.3 *Let $\mathfrak{g}(A) = \mathfrak{g}$ be a generalized Kac-Moody algebra such that if α_i and α_j are two distinct imaginary simple roots then $(\alpha_i, \alpha_j) = a_{ij} < 0$. Let W denote the direct sum of the integrable highest weight $\mathfrak{g}_S^e$-modules $U(\mathfrak{n}_S^-) \cdot f_j$ for all $j \in I \backslash S$. Given $\lambda \in P_+$ such that $(\lambda, \alpha_i) > 0$ for every imaginary simple root α_i ($i \in I \backslash S$), the generalized Verma module $V^{L(\lambda)}$ has the form $T(W) \otimes L(\lambda)$, and is a standard (and in particular irreducible) $\mathfrak{g}^e$-module.*

Proof: Let $\lambda \in (\mathfrak{h}^e)^*$ be such that $(\lambda, \alpha_i) = \lambda(h_i) \neq 0$ for all $i \in S$. The generalized Verma module $V^{L(\lambda)} = U(\mathfrak{u}^-) \otimes L(\lambda)$ is a standard module by Proposition 4.2. Since $\mathfrak{u}^-$ is the free Lie algebra on $W = \coprod_k U(\mathfrak{n}_S^-) \cdot f_k$ (by Theorem 2.1), we conclude that $U(\mathfrak{u}^-) \otimes L(\lambda) \cong T(W) \otimes L(\lambda)$. $\square$

Corollary 4.4 *In the setting of Theorem 4.3 assume that $\lambda(h_i) = 0$ for all $i \in S$. Then tensor algebra $T(W)$ carries the natural structure of a standard $\mathfrak{g}^e$-module.*

Proof: The $\mathfrak{g}_S$-module $L(\lambda)$ is one-dimensional. $\square$

Note that Corollary 4.4 includes the case in [15], where $\mathfrak{g}_S$ is a Kac-Moody algebra and the generalized Kac-Moody algebra $\mathfrak{g}$ is formed by adjoining one additional imaginary simple root α_0 to $\mathfrak{g}_S$. This is done (see [15]) by taking $\mathfrak{g}$ to be the Lie algebra associated to the matrix:

$$\left(\begin{array}{c|c} (a_{ij})_{i,j \in S} & (a_{0j})_{j \in S} \\ \hline (a_{j0})_{j \in S} & a_{00} \end{array}\right)$$

where $a_{00} \leq 0$. The hypothesis of Theorem 4.3 is then satisfied, because the condition on the imaginary simple roots becomes vacuous. Let f be the generator in $\mathfrak{n}^-$ associated to the one imaginary simple root. Thus if W denotes the integrable highest weight module $U(\mathfrak{n}_S^-) \cdot f$, then $T(W)$ is a standard module. It is isomorphic to the generalized Verma module associated to the module $\mathbb{C}_\lambda$ where $(\lambda, \alpha_0) = 1$ and $(\lambda, \alpha) = 0$ for simple $\alpha \in \Delta_R$.

Recall the matrix A given in Section 2.3, so that the Lie algebra $\mathfrak{g}(A)$ has only only one real simple root and all of the $a_{ij} = (\alpha_i, \alpha_j) < 0$ for α_i, α_j imaginary simple, so the conditions of Corollary 4.4 are satisfied. In the case of

the Lie algebra $\mathfrak{g}(A)/\mathfrak{c}'$ of Proposition 2.4 and the Monster Lie algebra $\mathfrak{m}$, we can obtain standard modules by applying Corollary 4.4 to $\mathfrak{g}(A)$. If we choose $\lambda \in P_+$ satisfying the conditions

$$\begin{array}{rcll} \lambda(h_{-1}) & = & 0 & \\ \lambda(h_{ik}) & = & \lambda(h_{ij}) & \text{for each } i > 0,\ 1 \leq k, j \leq c(i), \end{array} \tag{15}$$

we get an irreducible generalized Verma module $V^{L(\lambda)}$ for $\mathfrak{g}(A)/\mathfrak{c}'$. To obtain $\mathfrak{m}$-modules, given any real numbers a, b such that $a > 0$ and $a + b > 0$ let $\lambda \in (h^e)^*$ be an element satisfying the conditions

$$\begin{array}{rcll} \lambda(h_{-1}) & = & 0 & \\ \lambda(h_{ik}) & = & \lambda(h_{ij}) = ai + b & \text{for each } i > 0,\ 1 \leq k, j \leq c(i). \end{array} \tag{16}$$

Then $\lambda \in P_+$, and the full center $\mathfrak{c}$ of $\mathfrak{g}(A)$ acts as zero on the module $V^{L(\lambda)}$. Thus $V^{L(\lambda)}$ is also an $\mathfrak{m}$-module.

Recall the M-module $\mathcal{V}$ given by (7), so that $\mathfrak{u}^- = L(\mathcal{V})$. We have shown that for the choices of λ given above the M-module $T(\mathcal{V}) = U(\mathfrak{u}^-)$ has a natural standard module structure for $\mathfrak{g}(A)/\mathfrak{c}'$ or for $\mathfrak{m}$.

Summarizing the above considerations, we describe $\mathfrak{m} = L(\mathcal{V}) \oplus \mathfrak{gl}_2 \oplus L(\mathcal{V}')$ (and $\mathfrak{g}(A)/\mathfrak{c}'$) in the following (obvious) way:

Theorem 4.5 *Let $\mathcal{V}$ be given by (7). The Lie algebra $\mathfrak{m}$ can be realized in a natural way as a Lie algebra of operators on the irreducible $\mathfrak{m}$-module $T(\mathcal{V})$, identified as above with a generalized Verma module, for any weight λ satisfying (16). The M-module structure on the tensor algebra $T(\mathcal{V})$ induced by the M-module structure on $\mathcal{V}$ is compatible with the action of M on $\mathfrak{m}$. The following operators generate $\mathfrak{m}$: $\mathcal{V}$ acting on $T(\mathcal{V})$ by left multiplication, $\mathfrak{gl}_2$ acting on $T(\mathcal{V})$ via the action induced by the $\mathfrak{gl}_2$-module structure of $\mathcal{V}$, and $\mathcal{V}'$ acting on $T(\mathcal{V})$ via its natural action on the induced module. Furthermore, these comments extend to the Lie algebra $\mathfrak{g}(A)/\mathfrak{c}'$ for weights λ satisfying (15).* $\square$

We suggest that these irreducible $\mathfrak{m}$-modules $T(\mathcal{V})$ (and similarly, $T(\mathcal{V}')$) are realizable in a conceptual way as the particular free associative algebras of operators described in our conjecture in Section 2.

5 Application to Borcherds' proof of the moonshine conjectures

We will apply the decomposition of $\mathfrak{m}$ as $\mathfrak{u}^+ \oplus \mathfrak{gl}_2 \oplus \mathfrak{u}^-$, where $\mathfrak{u}^+$ and $\mathfrak{u}^-$ are free Lie algebras, to simplify part of the proof appearing in [4] of the Conway-Norton conjectures [8]. It has already been seen [19] that the denominator identity for $\mathfrak{m}$ can be obtained with the help of the fact that $\mathfrak{u}^-$ is a free Lie algebra (but still using Borcherds' product formula for $j(q)$). This provides one simplification of

Borcherds' proof, because it requires less of the theory of Kac-Moody algebras to be generalized. Another simplification will be obtained here by computing the homology $H(\mathfrak{u}^-, \mathbb{C})$ of the subalgebra $\mathfrak{u}^-$, rather than computing the homology of the Lie algebra $\mathfrak{n}^+$ (or equivalently $\mathfrak{n}^-$) as in [4]. This avoids using results of [14] extended to generalized Kac-Moody Lie algebras.

We show that the replication formulas of [8] can be obtained from this homology result and calculations similar to those appearing in [6] for computing the dimensions of homogeneous subspaces of free Lie algebras.

5.1 Homology

It is easy to compute the homology of the free Lie algebra $L(V)$ for a vector space V: The following exact sequence is a $U(L(V)) = T(V)$-free resolution of the trivial module:

$$0 \to T(V) \otimes V \xrightarrow{\mu} T(V) \xrightarrow{\epsilon} \mathbb{C} \to 0 \tag{17}$$

where μ is the multiplication map and ϵ is the augmentation map. Note that if a group or a Lie algebra acts on V then the group or Lie algebra also acts on every term in the resolution (17) in a natural way, and therefore acts on $H(L(V), \mathbb{C})$.

One immediately observes from the resolution (17):

$$\begin{gathered} H_0(L(V), \mathbb{C}) = \mathbb{C} \\ H_1(L(V), \mathbb{C}) = V \cong L(V)/[L(V), L(V)] \\ H_n(L(V), \mathbb{C}) = 0 \text{ for } n \geq 2. \end{gathered} \tag{18}$$

(It is also well known how to compute $H_0(\mathfrak{a}, \mathbb{C})$ and $H_1(\mathfrak{a}, \mathbb{C})$ for an arbitrary Lie algebra $\mathfrak{a}$ [7].)

5.2 Adams operations

In this section p, q and t are commuting formal variables. The variables p^{-1} and q^{-1} will be used to index the $\mathbb{Z} \oplus \mathbb{Z}$-grading of our vector spaces. All of the M-modules we encounter are finite-dimensionally $\mathbb{Z} \oplus \mathbb{Z}$-graded with grading suitably truncated and will be identified with formal series in $R(M)[[p, q]]$. Definitions and results from [22] about the λ-ring $R(M)$ of finite-dimensional representations of M are applicable to formal series in $R(M)[[p, q]]$. We summarize the results of [1], [22] that we use below.

Let G be any finite group. The representation ring $R(G)$ is a λ-ring [22] with the λ operation given by exterior powers, so $\lambda^i V = \bigwedge^i V$ for $V \in R(G)$.

In the following discussion we let $W, V \in R(G)$. The operation $\bigwedge^i$ satisfies

$$\bigwedge^i (W \oplus V) = \sum_{n=0}^{i} \bigwedge^n (W) \otimes \bigwedge^{i-n} (V). \tag{19}$$

Define

$$\bigwedge_t(W) = \bigwedge^0(W) + \bigwedge^1(W)t + \bigwedge^2(W)t^2 + \cdots. \tag{20}$$

Then

$$\bigwedge_t(V \oplus W) = \bigwedge_t(V) \cdot \bigwedge_t(W). \tag{21}$$

The Adams operations $\Psi^k : R(G) \to R(G)$ are defined for $W \in R(G)$ by:

$$\tfrac{d}{dt} \log \bigwedge_t(W) = \sum_{n \geq 0} (-1)^n \Psi^{n+1}(W) t^n. \tag{22}$$

We use the following properties of the Ψ^k from [22]:

$$\Psi^k(V \otimes W) = \Psi^k(V) \otimes \Psi^k(W) \tag{23}$$

$$\Psi^k(V \oplus W) = \Psi^k(V) \oplus \Psi^k(W). \tag{24}$$

Thus the Ψ^k are ring homomorphisms. Also

$$\Psi^k \Psi^l(W) = \Psi^{kl}(W). \tag{25}$$

For a class function $f : G \to \mathbb{C}$, define

$$(\Psi^k f)(g) = f(g^k). \tag{26}$$

for all $g \in G$. Then if χ_V is a character of G,

$$\chi_{\Psi^k(V)}(g) = (\Psi^k \chi_V)(g) \tag{27}$$

(cf. [22]).

Now let W be a finite-dimensionally $\mathbb{Z} \oplus \mathbb{Z}$-graded representation of G such that $W_{(\gamma_1,\gamma_2)} = 0$ for $\gamma_1, \gamma_2 > 0$. We shall write

$$W = \sum_{(\gamma_1,\gamma_2) \in \mathbb{N}^2} W_{(-\gamma_1,-\gamma_2)} p^{\gamma_1} q^{\gamma_2}, \tag{28}$$

identifying the graded space and formal series. We extend the definition of Ψ^k to formal series $W \in R(G)[[p,q]]$ by defining $\Psi^k(p) = p^k$, $\Psi^k(q) = q^k$ and in general,

$$\Psi^k\Big(\sum_{(\gamma_1,\gamma_2) \in \mathbb{N}^2} W_{(-\gamma_1,-\gamma_2)} p^{\gamma_1} q^{\gamma_2}\Big) = \sum_{(\gamma_1,\gamma_2) \in \mathbb{N}^2} \Psi^k(W_{(-\gamma_1,-\gamma_2)}) p^{k\gamma_1} q^{k\gamma_2}. \tag{29}$$

Expressions involving $\bigwedge^i(W)$ for $W \in R(G)[[p,q]]$ are interpreted in the obvious way, so that (20) is meaningful and (19) and (21) are valid for our finite-dimensionally graded spaces. The identity (22) remains valid for $W \in R(G)[[p,q]]$ and the Ψ^k defined above by (29). Also, (23) - (25) remain valid.

5.3 Möbius inversion

Let G be a finite group and let W be a finite-dimensionally Γ-graded representation of G, where Γ is a finitely generated free abelian group. Then $W = \coprod_{\gamma\in\Gamma} W_\gamma$ where each W_γ is a finite-dimensional module for G. We prove a Möbius inversion formula involving functions $s, t : G \times \Gamma \to \mathbb{C}$ such that for fixed $\gamma \in \Gamma$, s and t are class functions of G. For each $g \in G$, s and t are functions taking Γ to $\mathbb{C}$; we sometimes write $s(\gamma)$, $t(\gamma)$, $\gamma \in \Gamma$. For example, we may take $t(g, \gamma) = Tr(g|W_\gamma)$, $g \in G$, $\gamma \in \Gamma$. If Γ has base $\{\gamma_1, \ldots, \gamma_k\}$, and if $\nu = a_1\gamma_1 + \cdots + a_k\gamma_k \in \Gamma$ ($a_i \in \mathbb{Z}$) and $d \in \mathbb{Z}$ then we say $d|\nu$ if $d|a_i$ for each i. For $\nu, \kappa \in \Gamma$ such that $\nu = d\kappa$ we write $\frac{\nu}{\kappa} = d$. Let μ be the Möbius function.

Lemma 5.1 *Let s and t be as above and let $\nu \in \Gamma$, $d \in \mathbb{Z}_+$. Then*

$$\sum_{d|\nu} \Psi^d t\left(\frac{\nu}{d}\right) = s(\nu) \tag{30}$$

if and only if

$$t(\nu) = \sum_{d|\nu} \mu(d)\Psi^d s\left(\frac{\nu}{d}\right). \tag{31}$$

Proof: We use a standard argument (cf. [6]). Assume equation (30). Then

$$\sum_{d|\nu} \mu(d)\Psi^d s\left(\frac{\nu}{d}\right) = \sum_{d|\nu} \mu(d)\Psi^d \left(\sum_{\delta|\frac{\nu}{d}} \Psi^\delta t\left(\frac{\nu}{d\delta}\right)\right)$$

$$\begin{aligned}
&= \sum_{d|\nu} \mu(d) \sum_{\delta|\frac{\nu}{d}} \Psi^{d\delta} t\left(\frac{\nu}{d\delta}\right) \\
&= \sum_{d\delta|\nu} \mu(d)\Psi^{d\delta} t\left(\frac{\nu}{d\delta}\right) \\
&= \sum_{\kappa|\nu} \sum_{d|\frac{\nu}{\kappa}} \mu(d)\Psi^{\nu/\kappa} t(\kappa) \\
&= \sum_{\kappa|\nu} \Psi^{\nu/\kappa} t(\kappa) \sum_{d|\frac{\nu}{\kappa}} \mu(d) = t(\nu).
\end{aligned}$$

That (31) implies (30) is similar. □

We will apply this inversion formula to the situation where $\Gamma = \mathbb{Z} \oplus \mathbb{Z}$ and $W \in R(G)[[p, q]]$ as in (28). For a pair $(i, j) \in \mathbb{Z} \oplus \mathbb{Z}$ our definition specializes to: $k|(i, j)$ if $k(m, n) = (i, j)$ for some $(m, n) \in \mathbb{Z} \oplus \mathbb{Z}$. For $(i, j) \in \mathbb{Z}_+ \oplus \mathbb{Z}_+$ we define

$$P(i, j) = \{a = (a_{rs})_{r,s\in\mathbb{Z}_+} \mid a_{rs} \in \mathbb{N}, \sum_{(r,s)\in\mathbb{Z}_+\oplus\mathbb{Z}_+} a_{rs}(r, s) = (i, j)\}.$$

We will use the notation $|a| = \sum a_{rs}$, $a! = \prod a_{rs}!$.

5.4 Replication

In this section we show that the formulas for the graded traces of elements of M acting on $V^\natural$ (i.e., the McKay-Thompson series) are replicable [8], in fact, we show that they are completely replicable in the sense of [27] (see also [9] and [10]). We also observe that they can be computed recursively, as in [4]. The free Lie algebra structure of $\mathfrak{u}^-$ allows us to perform a calculation analogous to one appearing in [6], computing the dimensions of homogeneous subspaces of a free Lie algebra. In this context the replication formulas occur quite naturally.

Recall the structure of $\mathfrak{u}^-$ and $\mathcal{V} = H_1(\mathfrak{u}^-)$ as $\mathbb{Z}\oplus\mathbb{Z}$-graded M-modules ((9), Lemma 2.2, Lemma 2.3, (18)). We index the grading by p^{-1} and q^{-1} as in (28); then write $\mathfrak{u}^-$ and $\mathcal{V} = H_1(\mathfrak{u}^-)$ as elements of $R[M][[p,q]]$:

$$\mathfrak{u}^- = \sum_{(m,n)} V^\natural_{mn} p^m q^n \tag{32}$$

and

$$\mathcal{V} = \sum_{(m,n)} V^\natural_{m+n-1} p^m q^n, \tag{33}$$

where here and below the sums are over all pairs (m,n) such that $m,n > 0$.

Define

$$H_t(\mathfrak{u}^-) = \sum_{i=0}^{\infty} H_i(\mathfrak{u}^-) t^i$$

and let $H(\mathfrak{u}^-)$ denote the alternating sum $H_t(\mathfrak{u}^-)|_{t=-1}$. Recall the Euler-Poincaré identity:

$$\bigwedge\nolimits_{-1}(\mathfrak{u}^-) = H(\mathfrak{u}^-). \tag{34}$$

Taking log of both sides of (34) results in the formal power series identity in $R(M)[[p,q]] \otimes \mathbb{Q}$:

$$\log \bigwedge\nolimits_{-1}(\mathfrak{u}^-) = \log H(\mathfrak{u}^-), \tag{35}$$

where we have

$$\log H(\mathfrak{u}^-) = \log(1 - H_1(\mathfrak{u}^-)) = -\sum_{n=1}^{\infty} \frac{1}{n} H_1(\mathfrak{u}^-)^n.$$

Formally integrating (22), with $W = \mathfrak{u}^-$, gives

$$\log \bigwedge\nolimits_t(\mathfrak{u}^-) = -\sum_{n\geq 0} \Psi^{n+1}(\mathfrak{u}^-) \frac{(-t)^{n+1}}{n+1}.$$

Then setting $t = -1$ gives:

$$-\log \bigwedge\nolimits_{-1}(\mathfrak{u}^-) = \sum_{k=1}^{\infty} \frac{1}{k} \Psi^k(\mathfrak{u}^-).$$

Since $H_1(\mathfrak{u}^-) = \mathcal{V}$, equation (35) and equations (32) and (33) imply

$$\sum_{k=1}^{\infty} \frac{1}{k} \Psi^k \Big(\sum_{(m,n)} V^{\natural}_{mn} p^m q^n \Big) = \sum_{k=1}^{\infty} \frac{1}{k} \Big(\sum_{(m,n)} V^{\natural}_{m+n-1} p^m q^n \Big)^k$$

$$\sum_{k=1}^{\infty} \sum_{(m,n)} \frac{1}{k} \Psi^k (V^{\natural}_{mn}) p^{mk} q^{nk} = \sum_{k=1}^{\infty} \frac{1}{k} \sum_{(i,j) \in \mathbb{Z}_+} \sum_{\substack{a \in P(i,j) \\ |a| = k}} \frac{|a|!}{a!} \prod_{r,s \in \mathbb{Z}_+} (V^{\natural}_{r+s-1})^{a_{rs}} p^i q^j$$

$$\sum_{(i,j)} \sum_{k | (i,j)} \frac{1}{k} \Psi^k (V^{\natural}_{ij/k^2}) p^i q^j = \sum_{(i,j)} \sum_{a \in P(i,j)} \frac{(|a|-1)!}{a!} \prod_{r,s \in \mathbb{Z}_+} (V^{\natural}_{r+s-1})^{a_{rs}} p^i q^j .$$

Take the trace of an element $g \in M$ on both sides of this formula, writing $c_g(m)$ for $\mathrm{Tr}(g|V_m)$:

$$\sum_{(i,j)} \sum_{k | (i,j)} \frac{1}{k} \Psi^k (c_g(ij/k^2)) p^i q^j = \sum_{(i,j)} \sum_{a \in P(i,j)} \frac{(|a|-1)!}{a!} \prod_{r,s \in \mathbb{Z}_+} c_g(r+s-1)^{a_{rs}} p^i q^j .$$

Equating the coefficients of $p^i q^j$ gives

$$H_{i,j} = \sum_{k | (i,j)} \frac{1}{k} H^{(k)}_{ij/k^2}, \tag{36}$$

where, using the notation of [27], we set (for each $g \in M$)

$$H^{(k)}_i = c_{g^k}(i)$$

and

$$H_{i,j} = \sum_{a \in P(i,j)} \frac{(|a|-1)!}{a!} \prod_{r,s \in \mathbb{Z}_+} c_g(r+s-1)^{a_{rs}} .$$

Formula (36) asserts precisely that the McKay-Thompson series

$$T_g(V^{\natural}) = \sum_{i \in \mathbb{Z}} c_g(i) p^i,$$

$g \in M$, satisfy the "replication formula" as it is written in [27] (see also [8], [9] and [10]).

Recall the definition of "replicability" given in [27]: that $H_{i,j} = H_{m,n}$ whenever $ij = mn$ and $\gcd(i,j) = \gcd(m,n)$. (See also [9], where this definition is related to the replication formulas of [8].) Since all of the "replicates" $T_{g^k}(V^{\natural})$ of the McKay-Thompson series $T_g(V^{\natural})$ are replicable, the $T_g(V^{\natural})$ are "completely replicable" ([27]; see also [10]). Thus we observe from (36):

Proposition 5.2 *The McKay-Thompson series* $T_g(V^\natural) = \sum_{i\in\mathbb{Z}} c_g(i)p^i$, $g \in M$, *are completely replicable functions.* □

The modular functions of [8] also satisfy (36), by [27], [23], [9] and [10].

In order to obtain recursion relations for the McKay-Thompson series, we apply Möbius inversion to (36), with the functions $t((i,j)) = (i+j)c_g(ij)$ and $s((i,j)) = (i+j)H_{i,j}$:

$$c_g(ij) = \sum_{\substack{k>0\\ k(m,n)=(i,j)}} \frac{1}{k}\mu(k)\Psi^k\left(\sum_{a\in P(m,n)} \frac{(|a|-1)!}{a!}\prod_{r,s\in\mathbb{Z}_+} c_g(r+s-1)^{a_{rs}}\right)$$

$$= \sum_{\substack{k>0\\ k(m,n)=(i,j)}} \frac{1}{k}\mu(k)\left(\sum_{a\in P(m,n)} \frac{(|a|-1)!}{a!}\prod_{r,s\in\mathbb{Z}_+} c_{g^k}(r+s-1)^{a_{rs}}\right). \qquad (37)$$

If Q is a polynomial expression in the $c_g(n)$, $g \in M$, $n \leq N$ we say that Q has *level* $\leq N$. Then (37) implies that

$$c_g(ij) = c_g(i+j-1) + \text{ an expression of level } \leq i+j-3. \qquad (38)$$

This observation is, of course, vacuous if i or $j = 1$. Taking $(i,j) = (2l,2)$ equation (38) gives

$$c_g(4l) = c_g(2l+1) + \text{ an expression of level } \leq 2l-1.$$

Taking $(i,j) = (2l+1,2)$ equation (38) gives

$$c_g(4l+2) = c_g(2l+2) + \text{ an expression of level } \leq 2l.$$

For $(i,j) = (l,4)$ equation (38) gives

$$c_g(4l) = c_g(l+3) + \text{ an expression of level } \leq l+1.$$

It follows that the value of $c_g(4l)$ is given by an expression of lower level whenever $4l > 2l+1$, i.e., for $l \geq 1$. Also, $c_g(4l+2)$ is given by an expression of lower level whenever $4l+2 > 2l+2$, i.e., for $l \geq 1$. Finally, equating the two expressions for $c_g(4l)$ shows that $c_g(2l+1)$ is given by an expression of lower level whenever $2l+1 > l+3$, i.e., for $l > 2$. Thus $c_g(n)$ is determined by expressions of lower level except when $n = 1,2,3,5$. Thus the values of the $c_g(n)$ are determined by the $c_h(1)$, $c_h(2)$, $c_h(3)$, $c_h(5)$, $h \in M$, and equation (37).

As in [4], we conclude that since both the McKay-Thompson series and the modular functions of [8] satisfy (37), all that is necessary to prove that these functions are the same (see [4]) is to check the initial data listed above.

References

[1] M. Atiyah, *K-Theory*, W.A. Benjamin, New York, 1967.

[2] R. Borcherds, Generalized Kac-Moody Lie algebras, *J. Algebra* **115** (1988), 501-512.

[3] R. Borcherds, Central extensions of generalized Kac-Moody Lie algebras, *J. Algebra* **140** (1991), 330-335.

[4] R. Borcherds, Monstrous moonshine and monstrous Lie superalgebras, *Invent. Math.* **109** (1992), 405-444.

[5] R. Borcherds, Sporadic groups and string theory, to appear.

[6] N. Bourbaki, *Lie Groups and Lie Algebras, Part 1*, Hermann, Paris, 1975.

[7] H. Cartan and S. Eilenberg, *Homological Algebra*, Princeton University Press, 1956.

[8] J. H. Conway and S. P. Norton, Monstrous Moonshine, *Bull. London Math. Soc.* **11** (1979), 308-339.

[9] C. J. Cummins and S. P. Norton, Rational Hauptmoduls are replicable, to appear.

[10] C. Ferenbaugh, Replication formulae for Monster-like functions, to appear.

[11] I. Frenkel, J. Lepowsky and A. Meurman, A natural representation of the Fischer-Griess Monster with the modular function J as character, *Proc. Natl. Acad. Sci. USA* **81** (1984), 3256-3260.

[12] I. Frenkel, J. Lepowsky and A. Meurman, *Vertex Operator Algebras and the Monster*, Academic Press, Boston, 1988.

[13] O. Gabber and V. Kac, On defining relations of certain infinite-dimensional Lie algebras, *Bull. Amer. Math. Soc.* **5** (1981), 185-189.

[14] H. Garland and J. Lepowsky, Lie algebra homology and the Macdonald-Kac formulas, *Invent. Math.* **34** (1976), 37-76.

[15] R. Gebert and J. Teschner, On the fundamental representation of Borcherds algebras with one imaginary simple root, *Lett. Math. Phys*, to appear.

[16] P. Goddard and C. Thorn, Compatability of the dual Pomeron with unitarity and the absence of ghosts in the dual resonance model, *Phys. Lett.* **43** No. 2 (1972), 235-238.

[17] K. Harada, M. Miyamoto and H. Yamada, A generalization of Kac-Moody Lie algebras, to appear.

[18] E. Jurisich, *Generalized Kac-Moody algebras and their relation to free Lie algebras*, Ph.D. thesis, Rutgers University, May, 1994.

[19] E. Jurisich, Generalized Kac-Moody Lie algebras, free Lie algebras and the structure of the Monster Lie algebra, *J. Pure and Applied Algebra*, to appear.

[20] V. Kac, Simple irreducible graded Lie algebras of finite growth, *Math. USSR Izvestija* **2** (1968), 1271-1311.

[21] V. Kac, *Infinite Dimensional Lie Algebras*, Cambridge University Press, third edition, 1990.

[22] D. Knutson, *λ-Rings and the Representation Theory of the Symmetric Group*, Lecture Notes in Mathematics **308**, Springer-Verlag, 1973.

[23] M. Koike, On replication formula and Hecke operators, preprint.

[24] J. Lepowsky, *Lectures on Kac-Moody Lie algebras*, Université Paris VI, 1978 (unpublished).

[25] J. Lepowsky, Generalized Verma modules, loop space cohomology and Macdonald-type identities, *Ann. Scient. Ec. Norm. Sup.*, 4 série **12** (1979), 169-234.

[26] R. Moody, A new class of Lie algebras, *J. Algebra* **10** (1968), 211-230.

[27] S. P. Norton, More on Moonshine, in: *Computational Group Theory*, Academic Press, London, 1984, 185-193.

[28] J.-P. Serre, *Lie Algebras and Lie Groups*, W.A. Benjamin, New York, 1965.

[29] J.-P. Serre, *Algèbres de Lie semi-simples complexes*, W.A. Benjamin, New York, 1966.

arXiv:alg-geom/9701013v1 26 Jan 1997

FAMILIES OF K3 SURFACES

RICHARD E. BORCHERDS[†], LUDMIL KATZARKOV, TONY PANTEV[‡], AND N. I. SHEPHERD-BARRON[*]

Contents

1. Introduction 1
2. Automorphic forms on moduli spaces of K3 surfaces 4
3. Some examples 8
References 9

1. Introduction

We will prove the following theorem and give some examples to show that most of the conditions in it are necessary. Recall that a family $X \to B$ of varieties over a base space B is called isotrivial if there is an étale covering $\tilde{B} \to B$ such that $X \times_B \tilde{B}$ is a trivial family over $\tilde{B}$.

Theorem 1.1. *Any complete family of minimal Kähler surfaces of Kodaira dimension 0 and constant Picard number is isotrivial.*

This is a generalization to surfaces of the well known fact that any complete family of complex elliptic curves is isotrivial, because any complex elliptic curve is automatically minimal, Kähler, of Kodaira dimension 0, and has Picard number 1. Roughly speaking, it gives some cases when moduli spaces of surfaces contain no complete subvarieties.

In this paper all varieties are smooth, and are defined over $\mathbb{C}$ except when we explicitly state otherwise.

Recall that any minimal Kähler surface of Kodaira dimension 0 is abelian, hyperelliptic, Enriques, or K3. We prove theorem 1.1 by treating these cases separately. We can quickly dispose of most of these cases and reduce to the case of projective K3 surfaces as follows.

We often want to deduce that a compact family of surfaces with extra structure (including at least a polarization) is isotrivial from the fact that a suitable moduli space is quasiaffine (so that all fibers in the family are

1991 *Mathematics Subject Classification.* Primary 14J28, 14J10; Secondary 11F55.
Key words and phrases. $K3$ surfaces, automorphic forms.
(†) Partially supported by Grant DMS-9401186 and by a Royal Society professorship.
(‡) Partially supported by NSF grant DMS-9500712.
(∗) Partially supported by a European Community HCM grant.

isomorphic as the image of any compact variety in a quasiaffine space must be a point). If we had a fine moduli space for the varieties in question this would be automatic, and in this case the family would be trivial and not just isotrivial. In general we only have a coarse moduli space. However, it is well known that, except for Enriques surfaces, this can be made fine by introducing a level n structure ($n \geq 3$) on $H^2(\cdot, \mathbb{Z})$. More precisely we must trivialize the local system whose fibers are $H^2(\cdot, \mathbb{Z}/3\mathbb{Z})$, and this can be done by taking a finite étale cover of the base space. This rigidification of $H^2(\cdot, \mathbb{Z}/3\mathbb{Z})$ provides a fine moduli space because any element of finite order of $GL_N(\mathbb{Z})$ which is the identity modulo $n \geq 3$ is the identity. (For Enriques surfaces we must instead rigidify the cohomology with coefficients in $\mathbb{Z}/3\mathbb{Z}$ of its K3 cover.) So to show that compact families are isotrivial we just have to show that the corresponding coarse moduli spaces are quasiaffine.

The moduli space of hyperelliptic surfaces is affine ([1]) so theorem 1.1 follows immediately for these surfaces because the moduli space cannot contain complete subvarieties of positive dimension, and hence all fibers in a family are isomorphic and hence the family is isotrivial by the remarks above. Similarly by theorem 1 of [3] the moduli space of Enriques surfaces is quasiaffine so theorem 1.1 again follows. For abelian surfaces theorem 1.1 follows from the case of K3 surfaces by looking at the associated Kummer variety and using the fact that the Picard number of a Kummer surface is 16 plus the Picard number of the abelian surface. For families of non-projective K3 surfaces, theorem 1.1 follows from a result of Fujiki [5, Theorem 4.8 (1)]. So theorem 1.1 follows from the case of projective K3 surfaces, which we treat at the end of section 2. The main idea of the proof is to construct automorphic forms on the period space of marked K3 surfaces with Picard lattice containing a given lattice, such that all zeros of the automorphic form correspond to K3 surfaces with a larger Picard lattice. Then we use the fact that the zeros of an automorphic form give an ample divisor on the moduli space.

We construct these automorphic forms using the denominator function of the fake monster Lie algebra described in [2], in much the same way that the denominator formula of the fake monster Lie superalgebra was used in [3] to prove that the moduli space of Enriques surfaces is quasiaffine. We can get the following more precise results by looking more carefully at the automorphic forms we construct. Recall that an S-K3 surface X for some Lorentzian lattice $S \subset II_{3,19}$ is a K3 surface with a fixed primitive embedding of S into the Picard group such that the image of S contains a semi-ample class. (A semi-ample class is a class D such that $D^2 > 0$ and $D.C \geq 0$ for all curves C on the K3 surface X.)

Theorem 1.2. *There is an automorphic form on the period space of marked S-K3 surfaces which vanishes only on divisors $t^\perp$ of vectors t in the dual T' of $T = S^\perp$ with $0 > (t,t) \geq -2$. (The period space is an open subset of the space of 2 dimensional positive definite subspaces of $T \otimes \mathbb{R}$, and $t^\perp$ is the set of subspaces in the period space orthogonal to t; see section 2.)*

By taking S to be a 1 dimensional lattice generated by a vector of norm 2 we get:

Theorem 1.3. *Any family $f : X \to B$ of smooth, polarized K3 surfaces with polarization of degree 2 over a projective variety B is isotrivial.*

If the period point of a marked S-K3 surface lies on the zero locus of the automorphic form of theorem 1.2 then the surface is S-bad in the following sense.

Definition 1.4. We will call an S-K3 surface S-bad if S is contained in a sublattice S_1 of the Picard lattice such that $\dim(S_1) = \dim(S) + 1$ and $|\det(S_1)| \leq 2|\det(S)|$.

Here are some examples of S-bad K3 surfaces. If there is a norm -2 vector in the Picard lattice orthogonal to S then the K3 surface is obviously S-bad (take S_1 to be the lattice generated by S and this vector). Any K3 surface whose period point lies on one of the zeros of the automorphic form of theorem 1.2 is also S-bad (take S_1 to be generated by S and a vector of the Picard lattice whose projection to T' is t). If S is generated by a vector of norm $2n$, so that S-K3 surfaces include K3 surfaces with a polarization of degree $2n$, then it is easy to check that a K3 surface is S-bad if its Picard lattice has a vector D of degree k such that $-2 \leq (D, D) - k^2/2n < 0$. (By adding a multiple of the polarization vector to D and possibly changing the sign of D we can also assume that $0 \leq k \leq n$.)

As an immediate consequence of theorem 1.2 we get:

Theorem 1.5. *Any non isotrivial family $f : X \to B$ of S-K3 surfaces over a projective variety B has at least one S-bad fiber.*

We now show why most of the conditions of theorem 1.1 are necessary.

If we take a fixed K3 surface of Picard number n and blow it up in a varying point we get a complete non isotrivial family of non minimal surfaces of Picard number $n+1$. Hence theorem 1.1 does not extend to non minimal surfaces.

We show that the Kähler condition is necessary. Take two distinct elliptic curves A and B and let B_A be the sheaf of germs of analytic maps from A to B. The sheaf B_A is a sheaf of abelian groups on A and $H^1(A, B_A)$ classifies all primary Kodaira surfaces with base A and fiber isomorphic to B ([1], V.5., p. 143-147). Represent B as a quotient $\mathbb{C}/\Lambda$. Choose a nonzero element $c \in H^2(A, \Lambda)$ and let $T \subset H^1(A, B_A)$ be the preimage of c under the natural map coming from the short exact sequence $0 \to \Lambda \to \mathcal{O}_A \to B_A \to 0$. The coset T is isomorphic to the quotient (stack) $H^1(A, \mathcal{O}_A)/H^1(A, \Lambda) = A/\mathbb{Z}^2$. By pulling back to A the universal family over T we obtain a complete non-isotrivial family of primary Kodaira surfaces. The latter are never Kähler since their first Betti number is equal to 3 and thus theorem 1.1 does not hold for non-Kähler surfaces.

The moduli space of curves of genus greater than 2 contains compact curves, so the analogue of theorem 1.1 for curves of high genus is false. By

taking products of these families with curves of genus 0,1, or greater than 1 we can find compact non isotrivial families of minimal surfaces of constant Picard number 2 with Kodaira dimensions $-\infty$, 1, or 2. So theorem 1.1 is not true for surfaces of nonzero Kodaira dimension.

Over fields of characteristic $p > 0$ there are complete non isotrivial families of supersingular K3 surfaces of constant Picard number 22. (See [12].) So theorem 1.1 is false in nonzero characteristics (with "Kähler" replaced by "projective"). However a plausible analogue of the characteristic zero result might be that in a complete nonisotrivial family of supersingular K3 surfaces the Artin invariant must jump.

In section 3 of the paper we construct a non isotrivial family of smooth polarized K3 surfaces. By theorem 1.1 the family cannot have constant Picard number and we find some explicit examples of S-bad fibers for some 1-dimensional S. This shows that the condition about constant Picard number in theorem 1.1 cannot be omitted in the case of K3 surfaces. (The hypothesis about constant Picard number is irrelevant for Enriques or hyperelliptic surfaces because all such surfaces have Picard numbers 10 and 2 respectively.)

The K3 surfaces of the family in section 3 all have a fixed point free involution with an Enriques surface as the quotient. We use theorem 1.1 to show that there is no global fixed point free involution of the whole family.

The only condition in theorem 1.1 that we have not shown is necessary is the restriction to 2 dimensional varieties. We do not know what happens in higher dimensions.

Acknowledgments: We would like to thank J. Jorgensen and A. Todorov for some interesting discussions about the results of this paper. In particular R. E. Borcherds would like to thank A. Todorov for explaining their idea of using automorphic forms to attack the Shafarevich conjecture for K3 surfaces; this led to the idea of using the automorphic forms of [B95] to prove results about moduli spaces of surfaces. Jorgenson and Todorov independently noticed that results about isotriviality of families of K3 surfaces would follow from the existence of automorphic forms with known zeros, and their preprint [8] has some overlap with the results of this paper, such as theorem 1.3. (The methods used in [8] to construct automorphic forms are quite different from the methods in this paper and are based on regularized determinants as in [7].) We would also like to thank the referee for pointing out some gaps.

2. Automorphic forms on moduli spaces of K3 surfaces

In this section we construct some automorphic forms with known zeros on certain period spaces of marked K3 surfaces with extra structure, which gives explicit examples of ample divisors on the corresponding moduli spaces since the set of zeros of an automorphic form is an ample divisor. The extra structure consists of a fixed primitive embedding of some lattice S of signature $(1, m)$ in the Picard lattice of the K3 surface. We call such

a K3 surface an S-K3 surface. We regard S as a fixed sublattice of the lattice $II_{3,19}$, and write T for the lattice $S^\perp$ of signature $(2, 19-m)$. The (hermitean) symmetric space of the lattice T is the set of norm 0 points in the complex projective space of $T \otimes \mathbb{C}$ whose real and imaginary parts span a 2 dimensional positive definite subspace of $T \otimes \mathbb{R}$. We recall from [4] that the moduli space of S-K3 surfaces can be identified with the quotient of the symmetric space of the lattice $T = S^\perp$ by some arithmetic group. The Baily-Borel theorem implies that the zero locus of an automorphic form is an ample divisor on a compactification of the moduli space, and hence the set of points of the quotient where an automorphic form does not vanish is a quasiaffine variety.

The main result of this section is a proof of theorem 1.2, which states that there is an automorphic form on the space of marked S-K3 surfaces which vanishes only on divisors of vectors $t \in T'$ with $0 > (t,t) \geq -2$. We will construct this automorphic form by first embedding the lattice T in the lattice $II_{2,26}$ and then restricting a certain automorphic form of weight 12 for this lattice to the symmetric space of T.

We can find a geometric interpretation of the K3 surfaces whose period points lie on the divisors in theorem 1.2 as follows. We know that the Picard lattice contains S. As $t \in T'$ and $II_{3,19}$ is unimodular we can find a vector $D \in II_{3,19}$ whose projection into T is t. Then the lattice $\langle S, D\rangle$ generated by S and D has the properties

- $\langle S, D\rangle$ has signature $(1, m+1)$
- $|\det(\langle S, D\rangle)| \leq 2|\det(S)|$.

because the projection of v into the orthogonal complement of S has norm of absolute value at most 2. Hence the Picard lattice of the K3 surface contains a lattice with the properties above. In particular any K3 surface for which there is a norm -2 vector in $S^\perp$ satisfies the condition above, as we can take D to be this norm -2 vector.

We now prove theorem 1.2.

Proof. We first construct some primitive embeddings of T into $II_{2,26}$. Corollary 1.12.3 of Nikulin [11] implies that we can primitively embed any lattice T into the unimodular lattice $II_{2,26}$ provided that $T \otimes \mathbb{R}$ embeds into $II_{2,26} \otimes \mathbb{R}$ and the minimum number of generators of T'/T is less than $\dim(II_{2,26}) - \dim(T)$. We can therefore find a primitive embedding of our lattice T into $II_{2,26}$ because the rank of the group T'/T is at most the dimension of S, so this rank plus the dimension of T is less than the dimension of $II_{2,26}$. We will write U for the orthogonal complement $T^\perp$ of T in $II_{2,26}$. Then T and U have the same determinant as T is a primitive sublattice of $II_{2,26}$.

We recall some properties of the function Φ defined in example 2 of section 10 of [2]. The properties of Φ we will use are that Φ is an automorphic form on the hermitian symmetric space of $II_{2,26}$ and its only zeros lie on the divisors of norm -2 vectors of $II_{2,26}$. Some other properties of Φ which we will not use are that its zeros all have multiplicity 1, it has weight 12, it is

the denominator function of the fake monster Lie algebra, its Fourier series is explicitly known, and it can be written explicitly as an infinite product.

The restriction of Φ to the hermitian symmetric space of T is an automorphic form, but will be identically 0 whenever U contains a norm -2 vector. We can get around this by first dividing Φ by a product of linear functions vanishing on the divisors of each of these norm -2 vectors before restricting it. This restriction is an automorphic form as in pages 200-201 of [2]. So in all cases we get an automorphic form Φ_T on the hermitian symmetric space of T whose only zeros lie on the hyperplanes of norm -2 vectors of $II_{2,26}$. Although we do not need it we can work out the weight of Φ_T as follows: the weight is increased by 1 each time we divide Φ by a linear function, so the final weight is the weight (=12) of Φ plus half the number of norm -2 vectors of U.

We would like to know the zeros of Φ_T in terms of vectors of T rather than in terms of norm -2 vectors r of the larger lattice $II_{2,26}$. These zeros correspond to the hyperplanes of the negative norm projections of the norm -2 vectors r of $II_{2,26}$ into T. The projection of r into U has norm at most 0 as U is negative definite, so the projection t of r into T is a vector of T' with $0 > (t,t) \geq -2$. This proves theorem 1.2.

For the period space of marked Enriques surfaces there is an automorphic form vanishing exactly on the points orthogonal to -2 vectors [3], so it is natural to ask if there is an automorphic form for polarized K3 surfaces vanishing exactly on the points orthogonal to a norm -2 vector in $S^\perp = T$, which would be much stronger than the result above. Theorem 1.3 says there is such a form for K3 surfaces with a polarization of degree 2, but Nikulin [10] has shown that no such form can exist for some large values of the polarization.

The zeros of the form in theorem 1.2 do not always have multiplicity one; in fact they often have some zeros of high multiplicity. We can work out the multiplicity of the zeros by counting numbers of vectors in the dual U' of the lattice U with given norm and given image in U'/U. (But notice that some hyperplanes can have higher multiplicity than one might expect because they get zeros from more than one vector t.)

We now give some examples for polarized K3 surfaces, so we take S to be a one dimensional lattice spanned by a primitive vector of norm $2n$ for some positive integer n. We can parameterize embeddings of T into $II_{2,26}$ by primitive norm $-2n$ vectors v in $-E_8$. To do this we simply identify $T = (-2n) \oplus (-E_8) \oplus (-E_8) \oplus H \oplus H$ with the sublattice $\mathbb{Z}v \oplus (-E_8) \oplus (-E_8) \oplus H \oplus H$ of $II_{2,26} = (-E_8) \oplus (-E_8) \oplus (-E_8) \oplus H \oplus H$. The lattice U is then the orthogonal complement of v in $-E_8$.

Here is a table of the number of norm -2 roots of U and the numbers of vectors a in the lattice U' of norm greater than -2 for values of $k = (a,v)$ between 0 and n.

$2n$	*roots*	$k=0$	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$	$k=7$
2	126	1	56						
4	84	1	64	14					
6	74	1	54	27	2				
8	126	1	0	56	0	1			
8	56	1	56	28	8	0			
10	60	1	44	33	12	1	0		
12	46	1	48	30	16	3	48	10	
14	44	1	42	35	14	7	0	21	2
14	72	1	28	27	27	1	1	27	0
$2n$	*roots*	$k=0$	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$	$k=7$

The numbers of vectors for other values of k can be worked out using the fact that this number does not change if k is replaced by $2n+k$ or by $-k$. For some values of n there is more than one line because the lattice E_8 can have several orbits of vectors of the same norm, corresponding to several different automorphic forms. The first line for $2n=8$ corresponds to a non primitive norm 8 vector of E_8 so does not correspond to a primitive embedding of T into $II_{2,26}$. Some of the entries are 0, corresponding to the fact that the automorphic forms do not always vanish on all the divisors of theorem 1.2 (so the divisor in theorem 1.2 is not necessarily a minimal ample divisor).

Example 2.1. We will work out exactly what the automorphic form for $2n=2$ looks like. Its weight is (weight of Φ)+(number of roots of $E_7)/2 = 12+126/2 = 75$. The zeros of this form come by taking $k=0$ or 1 in theorem 1.2. For $k=0$ we get a contribution of 1 to the multiplicity of the divisor of each norm -2 vector in T. For $k=1$ we get a contribution of 56 for each norm $-1/2$ vector in the dual T' of T. This does not mean that the automorphic form has zeros of multiplicity 56, because twice a norm $-1/2$ vector of T' is a norm -2 vector of T which has even inner product with all vectors of T. In particular the divisor of the norm -2 vectors of T is reducible: it has two components E_1 and E_2 corresponding to norm -2 vectors which have odd inner product with some vector of T and to norm -2 vectors which have even inner product with all vectors of T. The divisor E_1 is a zero of the automorphic form of multiplicity 1, and the divisor E_2 is a zero of multiplicity $1+56=57$, and these are all the zeros.

In particular this example proves theorem 1.3 of the introduction.

Example 2.2. Nikulin conjectured in [10] that there are only a finite number of lattices S such that there is an automorphic form vanishing only on S-K3 surfaces which have a norm -2 vector in $S^\perp$. We can find a few examples of lattices S with this property. Firstly if S is a unimodular Lorentzian lattice in L then $T=T'$ so S has this property. The unimodular Lorentzian lattices in L are $II_{1,1}$, $II_{1,9}$, and $II_{1,17}$. Secondly if S has determinant 2 and dimension 1 mod 8 then it has the property above as in example 1, so we also get the lattices (2), $(2)\oplus(-E_8)$, and $(2)\oplus(-E_8)\oplus(-E_8)$.

Example 2.3. If $2n$ is 4,6,8, or 10 then we can see from the table above that we can assume that either $(D,D) = -2, (D,P) = 0$ or $(D,D) = 0$. Hence if the period of the K3 surface is on a zero of the automorphic form then either the surface is singular (in the sense that the Picard group contains a -2 vector orthogonal to S) or its Picard lattice contains a nonzero element with zero self intersection number.

Lemma 2.1. *Any family of K3 surfaces with constant Picard number is, after a finite étale base change, a family of S-K3 surfaces for some Lorentzian lattice S.*

Proof. By the assumption about constant rank, and since the Picard group of a K3 surface is always a primitive sublattice of H^2, the Picard groups form a sub-local system of the local system of H^2's. Since the monodromy action on the Picard group is finite we can, after an étale base change, assume that this subsystem is constant. This means that there is a primitive embedding of a constant local system with fiber S into the local system of H^2's. By definition this is a family of S-K3 surfaces. This proves lemma 2.1.

We now prove theorem 1.1. By the remarks after theorem 1.1 we can assume that all the surfaces in the family are K3 surfaces. By lemma 2.1 we can assume that we have a family of S-K3 surfaces for some lattice S. As the K3 surfaces are projective, the lattice S is Lorentzian. By theorem 1.2 and the remarks near the beginning of this section, if the family is not isotrivial there must be surfaces whose Picard lattice is at least 1 more than the dimension of S. This contradicts the fact that all surfaces in the family have the same Picard number and proves theorem 1.1.

3. Some examples

In this section we construct an example of a complete non isotrivial family of smooth polarized K3 surface, to show that the hypothesis about the Picard number in theorem 1.1 cannot be left out.

Suppose that $A \to C$ is a complete one parameter family of principally polarized family of abelian surfaces. Such a family exists because the boundary of the Satake compactification of the moduli space has codimension $2 > 1$. Set $f : K = A/\{\pm 1\} \to C$. Since f is isotrivial in a neighborhood of its critical locus, we can simultaneously resolve the singularities via a map $\sigma : \tilde{K} \to K$ to get $\tilde{K} \to C$, a family of smooth K3 surfaces.

Let Θ be the relative principal polarization on A. It is well known that 2Θ is the pullback of a relatively ample divisor H on K. Let E denote the exceptional locus of σ. Then it is well known that on each geometric fiber E is uniquely even. So on the geometric generic fiber $\tilde{K}_{\bar{\eta}}$ there is a unique divisor class $\bar{L}$ such that $E_{\bar{\eta}} \sim 2\bar{L}$. Recall that the étale cohomology group $H^1_{et}(\cdot, \mathbb{G}_m)$ is the Picard group $Pic(\cdot)$, and $H^2_{et}(\cdot, \mathbb{G}_m)$ is the Brauer group $Br(\cdot)$. From the Hochschild-Serre spectral sequence

$$E_2^{pq} = H^p(Gal(\mathbb{C}(\bar{\eta})/\mathbb{C}(\eta)), H^q_{et}(\tilde{K}_{\bar{\eta}}, \mathbb{G}_m)) \Rightarrow H^{p+q}_{et}(\tilde{K}_{\eta}, \mathbb{G}_m)$$

we get an exact sequence

$$0 \to Pic(\tilde{K}_\eta) \to Pic(\tilde{K}_{\bar{\eta}})^{Gal(\mathbb{C}(\bar{\eta})/\mathbb{C}(\eta))} \to Br(\mathbb{C}(\eta)).$$

Since the Brauer group $Br(\mathbb{C}(\eta))$ of the function field of a complex curve is trivial (Tsen's theorem) it follows that E is even on the generic fiber $\tilde{K}_\eta$. It follows that $E \sim 2L + V$ for some L and for some V supported on fibers. By taking a ramified double cover of the base if necessary we can assume that V is even so that E is even. Say $E \sim 2M$. Put $B = 2\sigma^*(H) - M$, so that B is a polarization of degree 8 on $\tilde{K}$ provided that H is very ample on K, which is equivalent to the abelian surface A being indecomposable as a principally polarized abelian variety. (See [6], pages 773-787.) If, however, A is decomposable and therefore a product of elliptic curves then B is merely semi ample. Nevertheless take S to be the lattice generated by B. We have constructed a complete non isotrivial family of S-K3 surfaces. (The divisor B is not ample on every fiber; if we want this as well we can take the divisor class $D = B + \sigma^*(H)$ which provides a polarization of degree $(3H)^2 + M^2 = 3^2 \times 4 - 8 = 28$.)

Example 3.1. Suppose that $0 \in C$ is such that A_0 is isomorphic to $X \times X'$ for elliptic curves X and X'. Then there is a morphism $\alpha : K \to \mathbb{P}^1 \times \mathbb{P}^1$ of degree 2, since $\mathbb{P}^1$ is the Kummer variety of an elliptic curve. Let F_1, F_2 be fibers of the projections $\mathbb{P}^1 \times \mathbb{P}^1 \to \mathbb{P}^1$. Let D_i be the pullback of F_i to $\tilde{K}_0$. We know that $(D_i, D_i) = 0$ as D_i is a fiber of a morphism onto a curve. Moreover $(D_i, E_0) = 0$, where E_0 is the exceptional locus on $\tilde{K}_0$. Since $B = 2(D_1 + D_2) - E/2$ it follows that $(B, D_i) = 4$, so that the lattice generated by $S = \langle B \rangle$ and D_1 has discriminant -16. This is an example of an S-bad fiber.

Remark. Each of the K3 surfaces in the family above is a Kummer surface and so by a result of [9] has a fixed point free involution (not necessarily unique!) such that the quotient is an Enriques surface. But by theorem 1.1 we cannot find a global fixed point free involution acting on the whole family, otherwise we would get a complete nonisotrivial family of Enriques surfaces.

References

[1] **W. Barth, C. Peters, A. Van de Ven** "Compact complex surfaces", Springer Verlag, 1984.

[2] **R. E. Borcherds** *Automorphic forms on $O_{s+2,2}(R)$ and infinite products*, Inventiones Mathematicae, 120, 1995, pp. 161-213.

[3] **R. E. Borcherds** *The moduli space of Enriques surfaces and the fake monster Lie superalgebra*, Topology vol. 35 no. 3, 699-710, 1996.

[4] *Geometrie des surfaces K3: Modules et periodes*, Astérisque, 126, 1985.

[5] **A.Fujiki,** *On primitively symplectic compact Kähler V-manifolds of dimension four.* in Classification of algebraic and analytic manifolds, (Katata, 1982), 71-250, Progr. Math., 39, Birkhäuser Boston, Boston, Mass., 1983.

[6] **P. Griffiths, J. Harris** "Principles of algebraic geometry", Wiley, 1978.

[7] **J. Jorgenson, A. Todorov** An analytic discriminant for polarized algebraic K3 surfaces, 1994 preprint.

[8] **J. Jorgenson, A. Todorov** Ample divisors, automorphic forms, and Shafarevich's conjecture, 1996 preprint.

[9] **Jong Hae Keum** *Every algebraic Kummer surface is the $K3$-cover of an Enriques surface*, Nagoya Math. Journal 118 (1990), pp. 99–110.

[10] **V. Nikulin** *The remark on discriminants of K3 surfaces moduli as sets of zeros of automorphic forms*, Preprint alg-geom/9512018.

[11] **V. Nikulin** *Integer symmetric bilinear forms and some of their geometric applications*, Izv. Acad. Nauk SSSR, Ser. Math. 43 (1979), no. 1, 111-177, 238. English translation in Mathematics of the USSR Izvestia, Vol. 14 No. 1 1980, 103-167.

[12] **A. N. Rudakov, T. Zink, I. R. Shafarevich.** *The influence of height on degenerations of algebraic surfaces of type K3*, Izv. Acad. Nauk SSSR, Ser Math 46, 117-134 (1982), translation in Math. USSR, Izv. 20 No. 1, 119-135 (1983). Also reprinted in "Collected mathematical papers" by I. R. Shafarevich, Springer 1989.

Mathematics department, Evans Hall, 3840, UC Berkeley, CA 94720-3840
Current address: D.P.M.M.S., 16 Mill Lane, Cambridge CB2 1SB, UK
E-mail address: reb@pmms.cam.ac.uk

M.S.R.I., 1000 Centennial Drive, Berkeley, CA 94720
Current address: Department of Mathematics, UC Irvine, Irvine, CA 92697-3875
E-mail address: lkatzark@math.uci.edu

Department of Mathematics, MIT, Cambridge, MA 02139
E-mail address: pantev@math.mit.edu

D.P.M.M.S., 16 Mill Lane, Cambridge CB2 1SB, UK
E-mail address: nisb@pmms.cam.ac.uk

Vertex algebras. *10 June 1997, corrected 19 June, 26 June, 26 September.*

Richard E. Borcherds, *
D.P.M.M.S., 16 Mill Lane, Cambridge, CB2 1SB, England.

arXiv:q-alg/9706008v2 22 Sep 1997

Contents.

1. Introduction.
2. Formal groups and Hopf algebras.
3. Vertex groups
4. Relaxed multilinear categories
5. The representations of a vertex group.
6. G vertex algebras
7. Vertex operators
8. Free field G vertex algebras.
9. The main identity
10. The Yang-Baxter equation
11. Cohomology of G vertex algebras.
12. Relation to quantum field theory.

1. Introduction.

In this paper we try to define the higher dimensional analogues of vertex algebras. In other words we define algebras which we hope have the same relation to higher dimensional quantum field theories that vertex algebras have to one dimensional quantum field theories (or to "chiral halves" of two dimensional conformal field theories).

The ideas in this paper are not really in final form. On the other hand, I have been rewriting versions of this paper for about 12 years, and it has by now become obvious that it never will attain a satisfactory final form. I apologize in advance for the resulting chaotic and incomplete nature of some of the sections.

The main ideas of this paper are as follows. We first define vertex groups, which can be thought of roughly as groups together with some allowable singularities for functions on the group. We look at the category of representations of the group, and redefine multilinear maps between representations to allow maps with certain sorts of singularities. The result of this is that the composition of multilinear maps is no longer a multilinear map, but it is sufficiently close to a multilinear map that we can still sensibly define analogues of associativity, G-algebras, commutative G-algebras, and so on. We will refer to the analogues of G algebraic structures in this relaxed multilinear category as G vertex algebraic structures; for example we can define G vertex (associative) algebras, G vertex Lie algebras, and so on. We will mostly be interested in commutative G vertex algebras, which are the analogues of commutative rings acted on by the "group" G. We show that vertex algebras are exactly the same as commutative associative G vertex algebras, where G is the simplest nontrivial example of a vertex group. It is now obvious how to define

* Supported by NSF grant DMS-9401186 and a Royal Society research professorship.

higher dimensional analogues of vertex algebras: they are just the commutative associative G vertex algebras for a higher dimensional vertex group G.

Many examples of commutative G vertex algebras come from quantum field theories. Very roughly, a quantum field theory ought to give a G algebra if and only if it has "good" operator product expansions, and the quantum field theory is then determined by its G vertex algebra. The underlying group of G will be the group of translations of spacetime, or some larger group such as the Poincare group, and the "allowable" singularities on G are the singularities that appear in the operator product expansions. Any G vertex algebra automatically has "good" operator product expansions; in fact, G vertex algebras are just a formalization of what physicists do with operator product expansions. A simple example of a higher dimensional commutative G vertex algebra is the space of Wick polynomials in a free field, which is the G vertex algebra of a free quantum field. The commutativity of the G algebra is of course very closely related to the "locality" property of a quantum field theory. On the other hand not all G vertex algebras come from quantum field theories, because there is usually no sensible way to construct a Hilbert space from a G vertex algebra.

G vertex algebras are also closely related to operator product expansion (OPE) algebras. The latter do not seem to be rigorously defined anywhere, and G vertex algebras can be thought of as an attempt to give a rigorous definition.

The simplest examples of G vertex algebras are associative algebras acted on by the underlying group of the vertex group G. In general a G vertex algebra can be thought of as a sort of associative algebra acted on by a group G, except that the ring multiplication has singularities and is not defined everywhere. G vertex algebras have most of the formal properties of associative algebras: for example, we can define left and right ideals, homomorphisms, tensor products, commutative G vertex algebras, and so on. We can also define left and right modules over G vertex algebras and multilinear maps between them, and these multilinear maps are sometimes represented by "tensor products". From this viewpoint quantization can be described as follows: quantization means deforming an honest commutative algebra acted on by the underlying group of G into a commutative G vertex algebra over a vertex group G. The commutative algebra we start with will usually be an associative algebra generated by classical fields and their derivatives (of all orders). The correlation functions of the quantized theory can be recovered from the G vertex algebra if we are given a trace on the G vertex algebra.

Another way to think of G vertex algebras is as follows. If V is an associative ring acted on by a group G, then for any fixed $v_i \in V$ and any $g_i \in G$ we can think of $v_1^{g_1} v_2^{g_2} \cdots v_i^{g_i}$ as being a function on G^i with values in V, and it is not difficult to write down a set of axioms for these functions equivalent to the axioms for an associative algebra acted on by G. The definition of a G vertex algebra is now in principle very easy: we use these axioms, except that we allow the functions from G^i to V to have some sort of singularities.

In practice it is necessary to allow G to be slightly more general than a group: for example it could be a formal group or Lie algebra. To include all these cases together we use cocommutative Hopf algebras H. A typical example of a cocommutative Hopf algebra is the group ring of a group, and representations of the group are the same as representations of its group ring. Similarly formal groups and Lie algebras can also be

considered as special cases of Hopf algebras. In most of this paper G will not be a group, but will be a cocommutative Hopf algebra together with a "vertex structure", as described below.

Informally we can think of any Hopf algebra H as being something like the group ring of a group, and can think of its dual H^* as being something like the ring of regular functions on this group. In order to define G vertex algebras we will need to know what is meant by a "singular function on G", and to specify these we need to give a vertex structure on the Hopf algebra H. This means that we are given an algebra K over the ring H^* which behaves as if it were the ring of singular functions on some group G.

The simplest nontrivial example of a vertex structure on a Hopf algebra is given by taking the Hopf algebra H to be the Hopf algebra of the one dimensional formal group, and by defining the ring of singular functions to be the quotient field of the ring of regular functions. Then the commutative G vertex algebras for this G turn out to be exactly the same as vertex algebras (as defined in [B] or [K]). (The (possibly non-commutative) G vertex algebras for this G are also equivalent to previously defined algebraic structures: they are more or less the same as "field algebras" defined in [K] and to "quantum algebras" defined in [L-Z].) In other words vertex algebras are just commutative G vertex algebras for the simplest possible nontrivial example G of a Hopf algebra with a vertex structure.

It seems to be hard to construct interesting examples of G vertex algebras for higher dimensional G, possible because they tend to be closely related to higher dimensional quantum field theories, which are notoriously difficult to construct. However we can construct G vertex algebras corresponding to generalized free field theories, which we do in section 8.

In section 9 we prove the main theorem of this paper, which is a generalization of the vertex algebra identity [K, 4.8] to G vertex algebras for higher dimensional G. The main result is that (roughly speaking) the integral of a vertex operator over an n-dimensional cycle is a vertex differential operator of order n.

Section 10 and 11 give a few miscellaneous thoughts about G vertex algebras.

There have been several previous definitions of mathematical structures inspired by quantum field theory; in particular there are various rigorous definitions of a quantum field theory based on operators on Hilbert spaces (see [S-W] or [G-J] for example), and there is Atiyah and Segal's notion of a topological field theory [Se]. In section 12 we discuss the relation between G vertex algebras and these other concepts.

Beilinson and Drinfeld [B-D] have another approach to defining vertex algebras, as Lie algebra objects in a suitable category. They remark that their definition extends to higher dimensions, but the main emphasis is on the extension to higher genus curves. I do not know what the relation of this is to this paper. The pseudo tensor categories in [B-D] are essentially the same as the multilinear categories of this paper. Soibelman [So] has some unpublished notes which overlap with the present paper. In particular he defines "meromorphic pseudo tensor categories" which are very similar to relaxed multilinear categories, and considers associative algebras in them. His notes include a generalization to the "braided" case, which we largely ignore in this paper.

This is based on a lecture that I promised but did not give at the 1996 Taniguchi symposium. I would like to thank the Taniguchi foundation for inviting me to Japan,

and C. Snydal and N. R. Scheithauer and the referee for suggesting many corrections and improvements.

2. Formal groups and Hopf algebras.

As well as considering groups acting on modules, we also need to consider various group-like things such as Lie algebras or formal groups acting on spaces. The most convenient way of doing these is to use cocommutative Hopf algebras, which are a common generalization of groups, Lie algebras, and formal groups. In this section we recall some standard results about formal groups and Hopf algebras, most of which can be found in any standard references such as [A] for Hopf algebras and [S] for formal groups.

As motivation for the definition of Hopf algebras we consider the following example.

Example 2.1. Suppose G is any discrete group and H is its group ring over some commutative ring R. Then H is an associative algebra with identity, and also has a map Δ (called the comultiplication) from H to $H \otimes_R H$ given by $\Delta(g) = g \otimes g$ for $g \in G$, and a map ϵ (called the counit) from H to R given by $\epsilon(g) = 1$, and a map S (called the antipode) from H to H given by $S(g) = g^{-1}$. Then H has the following properties:

1. H is an associative algebra with identity over R.
2. Δ is an algebra homomorphism from H to $H \otimes H$, which is coassociative, and ϵ is a homomorphism of algebras from H to R, which is a counit for the comultiplication.
3. $\mu(S \otimes 1(\Delta(h))) = \mu(1 \otimes S(\Delta(h))) = \epsilon(h)$ for all h.

Definition 2.2. *A* **Hopf algebra** *H is defined to be a module over R with multiplication, comultiplication, identity, counit, and antipode S, satisfying the axioms above. If in addition the comultiplication is cocommutative then the Hopf algebra is called* **cocommutative**.

So the group ring of any group is a cocommutative Hopf algebra. Conversely any cocommutative Hopf algebra behaves in many ways as if it were the group ring of some group.

The comultiplication is rather difficult to handle in notation. We assume that for each $h \in H$ we have chosen finite sets of elements h_{1i} and h_{2i} such that $\Delta(h) = \sum_i h_{1i} \otimes h_{2i}$. If H is a Hopf algebra then the comultiplication on H induces a multiplication on the dual $H^* = \mathrm{Hom}_R(H, R)$ which makes H^* into an associative algebra over R with unit η. If H is cocommutative then H^* is commutative. If H is a finitely generated free module over R then H^* also has a coalgebra structure making it into a Hopf algebra induced by the product of H, but in general H^* does not have a coalgebra structure, because the natural map from $H^* \otimes_R H^*$ to $(H \otimes_R H)^*$ is not usually an isomorphism.

Example 2.3. If L is a Lie algebra over R then the universal enveloping algebra $U(L)$ of L is a Hopf algebra, with the comultiplication induced by $\Delta(g) = g \otimes 1 + 1 \otimes g$ and S induced by $S(g) = -g$ for $g \in L \subset U(L)$. If R is a field of characteristic 0 then Lie algebras over R are essentially equivalent to formal groups, and the universal enveloping algebra of g is just isomorphic to the Hopf algebra of the formal group corresponding to g (defined below). If R is not an algebra over $\mathbf{Q}$ then the Hopf algebra $U(L)$ usually does not have particularly good properties, and the Hopf algebras of formal groups behave much better (at least for the purposes of this paper).

A formal group of dimension n over a commutative ring R (see [S, II.4.6]) is an n-tuple $F(x,y) = (F_1(x,y), \dots, F_n(x,y))$ of formal power series in $2n$ variables $x = (x_1, \dots, x_n), y = (y_1, \dots, y_n)$ such that

(1) $F(0,y) = y$, $F(x,0) = x$.

(2) $F(x, F(y,z)) = F(F(x,y), z)$.

Axiom 1 implies the existence of unique power series $\phi(x) = (\phi_1(x), \dots, \phi_n(x))$ with $\phi(0) = 0$ and

(3) $F(x, \phi(x)) = 0 = F(\phi(x), x)$.

We will abbreviate $F(x,y)$ to xy and $\phi(x)$ to x^{-1}, so that the axioms above become the same as the usual group axioms (except that the multiplicative identity is rather confusingly written as 0).

Example 2.4. The formal group of GL_n. This formal group has n^2 variables $x = (x_{ij})$ which we think of as the n^2 entries of an $n \times n$ matrix. The power series F and ϕ are defined by $I_n + F(x,y) = (I_n + x)(I_n + y)$ and $I_n + \phi(x) = (I_n + x)^{-1}$ where I_n is the identity matrix. So for example $F_{ik}(x,y) = x_{ik} + y_{ik} + \sum_j x_{ij} y_{jk}$.

We can also define formal groups of countable infinite dimension in a countable infinite number of variables x_i, $i \in N$, provided we are careful to use the correct definition of formal power series in an infinite number of variables: a formal power series in variables x_i is defined as a formal infinite sum $\sum_{0 < i_1 \le i_2 \le \dots \le i_n} a_{i_1 i_2 \dots i_n} x_{i_1} x_{i_2} \cdots x_{i_n}$. In particular the ring of formal power series in an infinite number of variables is strictly larger than the completion of the ring of polynomials in the x_i's at the ideal generated by the x_i's, because in the latter there would only be a finite number of nonzero a's for any given value of n.

Example 2.5. Suppose G is a formal group of dimension n. We can construct a cocommutative Hopf algebra RG, called the formal group ring of G, as follows. We let RG be a free module over R with a basis of elements $D^{(i)}$, where $i = (i_1, \dots, i_n)$ is an n-tuple of nonnegative integers. (If the dimension n is infinite we should add the condition that all but a finite number of the i_j's are zero.) We define the comultiplication by $\Delta(D^{(i)}) = \sum_j D^{(j)} \otimes D^{(i-j)}$, and the antipode S by $S(D^{(i)}) = (-1)^{i_1 + \dots i_n} D^{(i)}$. We can identify RG with the topological dual of the ring of formal power series $R[[x]]$ by letting $D^{(i)}$ be a dual basis to the elements $x^i = x_1^{i_1} \cdots x_n^{i_n}$ in the obvious way. Then the continuous algebra homomorphism from $R[[x]]$ to $R[[x,y]]$ taking x to $F(x,y)$ induces a dual map from $RG \otimes RG$ to RG which we use as the algebra structure on RG. (Alternatively we could also define RG to be the algebra of "continuous left invariant differential operators on G".) If we work over a field of characteristic 0, then formal groups are essentially the same as Lie algebras, and the formal group ring of a formal group is isomorphic to the universal enveloping algebra of the corresponding Lie algebra.

Example 2.6. Suppose G_a is the one dimensional formal group with formal group law $F(x,y) = x + y$. Then the formal group ring H over R is the R-algebra generated by elements $D^{(i)}$ for i a non-negative integer, with $D^{(0)} = 1$, $D^{(i)} D^{(j)} = \binom{i+j}{i} D^{(i+j)}$. If R contains the rational numbers this is just the ring of polynomials in one variable $D = D^{(1)}$, with $D^{(i)} = D^i / i!$. The dual H^* can be identified with the ring of Laurent power series $R[[x]]$, in such a way that the elements x^i are a dual "basis" to the elements $D^{(i)}$.

The category of (left) modules over a Hopf algebra H is defined to be the category of left modules over H considered as an R-algebra. (If A is a right module over R we can

turn it into a left module by defining $ha = aS(h)$.) If A and B are modules over H then $A \otimes_R B$ and $\mathrm{Hom}_R(A, B)$ can be made into H-modules in a canonical way as follows.

$$h(a \otimes b) = \Delta(h)(a \otimes b)$$

$$(hf)(a) = \sum h_{(1)}(f(S(h_{(2)})(a)))$$

In particular the modules over a Hopf algebra form a tensor abelian category.

If P is any free module over R then the tensor algebra $T^*(P) = \oplus_{n\geq 0} \otimes^n (P)$ has a natural coalgebra structure, given by $\Delta(p_1 \otimes \cdots \otimes p_n) = \sum_{0\leq j\leq n}(p_1 \otimes \cdots \otimes p_j) \otimes (p_{j+1} \otimes \cdots \otimes p_n)$. We denote the largest cocommutative subcoalgebra of this by $B(P)$, so that $B(P) = \oplus_{n\geq 0}(\otimes^n P)^{S_n}$, where $(\otimes^n P)^{S_n}$ is the subspace of $(\otimes^n P)$ fixed by the natural action of the symmetric group S_n. Over a field of characteristic 0 the natural map from $B(P)$ to the symmetric algebra of P is an isomorphism of vector spaces, but this is not true in positive characteristic. Warning: the symmetric coalgebra $B(P)$ is not the universal cocommutative coalgebra cogenerated by P (although some books incorrectly state that it is, probably because of the analogy with the symmetric algebra). The relation between them is that $B(P)$ is an irreducible component of the universal cocommutative coalgebra cogenerated by P; see [A]. We can recover P from $B(P)$ as the space of primitive elements, i.e., the elements p with $\Delta(p) = p \otimes 1 + 1 \otimes p$. A Hopf algebra is the formal group ring of a formal group over R if and only if it is a free R module and its underlying coalgebra is isomorphic to $B(P)$ for some free R module P.

3. Vertex groups.

In this section we define the concept of a vertex group G, which can be thought of informally as a group with a space of functions with singularities on it. We will later define G vertex algebras for any vertex group G as analogues of algebras acted on by an ordinary group G. The definition of a vertex group has to be weak enough to include the examples later, and strong enough so that it is possible to define the concept of an associative algebra over a vertex group.

In the definition below it is helpful to think of H as the Hopf algebra given by the group ring of some group G, so that H^* can be thought of as something like the ring of functions on the underlying space of G, and the left and right actions of H on H^* correspond to the left and right actions of G on itself. The module K below should then be thought of as some sort of ring of rational or algebraic functions with singularities on G. The axioms for a vertex structure are based on obvious properties of the space of singular functions on a group. In particular the singular functions form an algebra over the regular functions on which we can define the operations of left and right translation by elements of the group and the operation induced by taking inverses of elements of the group.

Definition 3.1. *Suppose that H is a Hopf algebra over a commutative ring R. An* **elementary vertex structure** *on H consists of an R-module K with the following extra structures.*

(1) K has the structure of an associative algebra over H^*.

(2) K has the structure of a 2 sided H-module, and the natural map from H^* to K is a homomorphism of 2 sided H-modules. The product on K is invariant under the left

and right actions of H. This means that $h(ab) = \sum (h_{(1)}a)(h_{(2)}b)$ and similarly for the right action. This axiom can be thought of as saying that K is closed under "left and right translation".

(3) There is an R linear map S (called the antipode) from K to K extending the antipode S on H^*, and $S(ab) = S(b)S(a)$ whenever each of a and b is in H or H^* or K. This axiom can be thought of as saying that K is closed under the inversion map of G.

Definition 3.2. *An* **(elementary) vertex group** *G is a Hopf algebra H with an (elementary) vertex structure K, such that H is cocommutative, K is commutative, and $S^2 = 1$. We call H the group ring of the vertex group G, and we call K the ring of singular functions on G.*

If we replace the condition $S^2 = 1$ by the weaker condition that S is an isomorphism of R modules from K to K then we call G a **braided elementary vertex group**. There are several other obvious variations of these definitions (which we will not use in this paper): if we add the condition that H is the Hopf algebra of a formal group we get the definitions of vertex formal groups and braided vertex formal groups, and if we replace the condition that the vertex structure is cocommutative by the condition that the Hopf algebra H is "braided quasi commutative" or "quasi triangulated" we get the concept of a "braided vertex quantum group".

The definition of an elementary vertex group is not really general enough for some examples. The main problem is that the only singularities of functions of several variables we allow are essentially singularities of functions of two variables, and there are some examples where we want to allow more general singularities; see the end of section 8 for examples and a provisional definition of a non elementary vertex group. However the definition above is adequate for most of the examples in this paper (possibly because most of the examples are related to quantum field theories that are either free or in small dimensions). We will usually drop the adjective "elementary" from now on.

Example 3.3. If H is any cocommutative Hopf algebra, then taking $K = H$ gives H the structure of a vertex group, called the trivial vertex group structure. In particular vertex groups are generalizations of cocommutative Hopf algebras (and hence of groups and Lie algebras).

Example 3.4. The simplest nontrivial example of a vertex group G is given as follows. Suppose that H is the Hopf algebra of the 1-dimensional formal group G_a of example 2.6, so that H has a basis of elements $D^{(i)}$ for $i \geq 0$ and H^* can be identified with $R[[x]]$ as in example 2.6. We let K be the quotient field $R[[x]][x^{-1}]$ of H^* consisting of formal Laurent series over R, with S acting as $S(x^i) = (-1)^i x^i$ and H acting as derivations in the obvious way. (The right and left actions of H on K are identical.) If G is the vertex group given by H and K then we will see later that (classical) vertex algebras are exactly the same as the commutative G vertex algebras as defined in section 6.

Example 3.5. Suppose the ring R contains the inverse $1/N$ of some integer N and has an element ζ with $\zeta^N = -1$. Then there is a variation of example 3.4 where we take K to be the ring $R[[x^{1/N}]][x^{-1/N}]$ of all formal Laurent series in $x^{1/N}$, with S acting as $S(x^{1/N}) = \zeta x^{1/N}$. This gives examples of braided vertex groups which are not vertex groups (because the antipode S does not have period 2).

Example 3.6. Take H to be the formal group of the algebraic group SL_2 (over any commutative ring R). If we represent an element of $SL_2(R)$ as $\binom{ab}{cd}$ then H^* is the ring of formal power series in $a-1, b, c, d-1$ modulo the ideal $(ad-bc-1)$. We can define a vertex structure by putting $K = H^*[b^{-1}]$, and this defines a vertex formal group. Under the same conditions as example 3.5 we can define a braided vertex formal group by putting $K = H^*[b^{-1/N}]$.

If R is an integral domain we can often define K to be the full quotient field of H^*, or even the separable algebraic closure of this quotient field, but this definition of K usually seems either too large or too small: for most of the examples later in this paper it is only necessary to invert a few elements of K and inverting more makes theorem 9.1 much weaker, and on the other hand we also sometimes want to allow transcendental extensions as in example 3.9 below.

Example 3.7. Take H to be the formal group of SL_n, so that H^* is the ring of formal power series in the elements $a_{ij} - \delta_i^j$ $(1 \le i, j \le n)$ modulo the ideal generated by $\det(a_{ij}) - 1$, where the a_{ij}'s are the entries in an n by n matrix. If D_k is the determinant of the top right $k \times k$ submatrix, so that $D_n = 1$ and $D_1 = a_{1n}$, then we can define a vertex structure by letting K be H^* with some subset of the D_k's inverted. (So example 3.6 is the case with $n = 2$ where we invert D_1.)

Example 3.8. We can generalize example 3.7 as follows. We let G be a split simple algebraic group with a fixed choice of Cartan subgroup T and Borel subgroup B. We define $\chi_1, \dots, \chi_n$ to be the fundamental characters of T in the Weyl chamber. These characters can be extended uniquely to functions on G which are right invariant under B and left invariant under $\bar{B}$, which we will again denote by $\chi_1, \dots, \chi_n$. For example if G is SL_n and T the diagonal matrices and B the upper triangular matrices then χ_i is given by the determinant of the top left $i \times i$ submatrix. Now take ω to be an element of G representing the opposition involution of the Weyl group, so that ω takes B to $\bar{B}$, and define functions D_k by $D_k(g) = \chi_k(\omega g)$, so that these functions are invariant under left and right multiplication by elements of B. We define a vertex structure by inverting some subset of the elements D_k, considered as formal power series in H^*. (Notice that changing ω only results multiplying D_k by an invertible power series, so that this vertex structure does not depend on the choice of representative ω.) Notice that the product of the functions D_k vanishes exactly on the complement of the big cell of the Bruhat decomposition, so that elements of K can be thought of informally as singular functions defined near the identity of G whose poles are in the complement of the big cell.

Example 3.9. Suppose that R is an algebra over the rational numbers and H is the Hopf algebra of the formal group of G_a as in example 3.4. Let K be the ring $R[[x]][x^{-1}, \log(x)]$ where $\log(x)$ is a new indeterminate (written as $\log(x)$ rather than y for mnemonic reasons) with $D^{(i)}(\log(x)) = (-1)^{i-1}x^{-i}/i$ if $i > 0$ and $S(\log(x)) = \log(x)$. (We could also define $S(\log(x)) = \log(x) + c$ for any $c \in R$ if we did not mind that $S^2 \ne 1$.) This defines a vertex structure on H which gives a vertex group. Similarly in examples 3.6 to 3.8 we could introduce formal logs of the functions b or D_k that we inverted.

We can define homomorphisms of vertex groups as follows. The definition is the obvious one if one thinks of a vertex group as a group ring.

Definition 3.10. *Suppose G_1 and G_2 are vertex groups with group rings H_1 and H_2 and*

rings of singular functions K_1 and K_2. Then a homomorphism of vertex groups from G_1 to G_2 consists of a homomorphism of Hopf algebras (over R) from H_1 to H_2, together with a H_2^*-algebra homomorphism from K_2 to K_1 which commutes with the antipode and the left and right actions of H_1.

Example 3.11. If G_1 and G_2 are "really" discrete groups, in other words if H_1 and H_2 are their group rings and $K_1 = H_1^*$, $K_2 = H_2^*$, then a homomorphism of vertex groups from G_1 to G_2 is the same as a homomorphism of groups from H_1 to H_2.

Example 3.12. Suppose G_1 is the trivial group (or rather the vertex group corresponding to it). If G_2 is also an honest group, then there is a unique homomorphism from G_1 to G_2. If G_2 is an arbitrary vertex group this is not usually true. For example if G_2 is the vertex group of example 3.4 with $K_2 = R[[x]][x^{-1}]$ then there is no homomorphism from G_1 to G_2, because the homomorphism from $H_2^* = R[[x]]$ to $H_1 = R$ cannot be extended to the field K_2. The significance of this is as follows. If G_1 and G_2 are groups then any homomorphism from G_1 to G_2 induces a forgetful functor from G_2 algebras to G_1 algebras in the obvious way. In particular if we take G_1 to be the trivial group this shows (in a rather roundabout way!) that any G_2 algebra has an underlying associative algebra structure. Similarly for G vertex algebras a homomorphism from G_1 to G_2 induces a natural forgetful functor from G_2 vertex algebras to G_1 vertex algebras. However we cannot use this to show that every G_2 vertex algebra has an underlying associative algebra structure, because the map from the trivial vertex group to G_2 need not exist.

Example 3.13. We can construct products of vertex groups with the usual universal properties. For example the product $G_1 \times G_2$ of two vertex groups G_1, G_2 is constructed as follows. The underlying Hopf algebra of $G_1 \times G_2$ is $H_1 \otimes H_2$, and the ring of singular functions is $(H_1 \otimes H_2)^* \otimes_{H_1^* \otimes H_2^*} K_1 \otimes K_2$.

Example 3.14. Suppose G_1 is the vertex group with $K_1 = R[[z]][z^{-1}]$ of example 3.4 and G_2 is the 2 dimensional vertex group $G_1 \times G_1$, so that $K_2 = R[[x,y]][x^{-1}, y^{-1}]$. For each $(a,b) \in R^2$ there is a homomorphism from H_1 to H_2 taking x to az and y to bz. This extends to a homomorphism of vertex groups if and only if $a \neq 0$ and $b \neq 0$.

4. Relaxed multilinear categories.

We can define multilinear maps between the representations of a vertex group, but the composition of multilinear maps is not in general multilinear. We deal with this problem by defining relaxed multilinear categories, where the composition of multilinear maps need not be multilinear, but can still be compared with multilinear maps. Symmetric relaxed multilinear categories have the following two properties: the representations of a vertex group form a symmetric relaxed multilinear category, and it is possible to define algebraic objects like commutative rings in a symmetric relaxed multilinear category. The reader willing to assume this can skip the rest of this section, which consists mainly of content-free category theory.

Soibelman has defined a similar notion in unpublished notes [So], except that his version is more general in several ways; for example, he allows braided rather than symmetric categories.

We will start by discussing what is needed in order to be able to define an associative algebra in some additive category. Obviously it is sufficient to assume the category is a

tensor category, in other words for every two objects there is a tensor product given with suitable properties. If we are given good isomorphisms between $A \otimes B$ and $B \otimes A$ for all A and B then we get the notions of symmetric and braided multilinear categories. In any additive symmetric tensor category we can define most common algebraic structures, such as associative algebras, commutative algebras, Lie algebras (ignoring problems in characteristic 2), Hopf algebras, and so on. If $A_1, \ldots, A_n, B$ are objects of any additive tensor category we have a space of multilinear maps from $A_1, \ldots, A_n$ to B, defined as the maps from $A_1 \otimes \cdots \otimes A_n$ to B.

We can weaken the definition of tensor category by assuming that we are just given spaces of multilinear maps from $A_1, \ldots, A_n$ to B for objects A_i and B, with suitable properties, but are not given objects $A_1 \otimes \cdots \otimes A_n$ representing them. (For example, composition is defined under obvious conditions, and is associative, and the identity maps behave in the obvious way.) An additive category with such a structure is called a multilinear category, and if we add conditions about the action of the symmetric group on spaces of multilinear maps we get the concepts of symmetric and braided multilinear categories. In a multilinear category we can still define associative algebras, commutative algebras, Lie algebras, and so on, but we cannot define things like coalgebras and Hopf algebras because these cannot be defined just in terms of multilinear maps but require maps to tensor products. For a multilinear category to be a tensor category it is necessary that spaces of multilinear maps should be representable, but this is not sufficient because there is no reason in general why the objects $(A \otimes B) \otimes C$, $A \otimes B \otimes C$, and $A \otimes (B \otimes C)$ should be isomorphic. A multilinear category with only one object is an operad.

A multilinear category is sometimes called an additive pseudo tensor category but this seems a rather misleading name: "tensor" suggests that tensor products exist, and "pseudo" then says that in fact they do not.

We now want to define "relaxed multilinear categories", which are a generalization of multilinear categories for which the composition of multilinear maps is not always a multilinear map. We start by defining the sieve category $Sieve_n$ as follows. Its objects are sieves of depth $k \geq 0$ on $1, 2, \ldots, n$, by which we mean a sequence of equivalence relations $E_0, E_1, \ldots, E_k$ on $1, \ldots, n$ such that

1 E_k is the "indiscrete" equivalence relation where any two elements are equivalent, and E_0 is the "discrete" equivalence relation where any element is only equivalent to itself

2 Any equivalence class in any equivalence relation is an interval $\{i, i+1, \ldots, j-1, j\}$

3 Any equivalence class of E_{i+1} is a union of equivalence classes of E_i (so that the equivalence relations are increasingly fine).

A sieve can be thought of as a record of someone's attempt to multiply n non-commuting elements of a ring. For example suppose someone tries to calculate the product $a_1a_2a_3a_4a_5a_6$. At the end of the first day they might have calculated $a_1a_2a_3$ and a_5a_6. At the end of the second day they might have worked out $a_1a_2a_3$ and $a_4a_5a_6$, and by the

third day they might have calculated $a_1a_2a_3a_4a_5a_6$. This would correspond to the sieve

$$E_0 = \{\{1\}\{2\}\{3\}\{4\}\{5\}\{6\}\}$$
$$E_1 = \{\{1,2,3\}\{4\}\{5,6\}\}$$
$$E_2 = \{\{1,2,3\}\{4,5,6\}\}$$
$$E_3 = \{\{1,2,3,4,5,6\}\}$$

There are two other useful ways of representing sieves of width n and depth d. The second is as some parentheses, such as ((((()())))((())(()()))), with the property that there are n innermost pairs of parentheses, each of which is contained in exactly d other pairs of parentheses, and such that there is an outermost pair containing everything. The pairs of parentheses correspond to equivalence classes of the equivalence relation, in a way that should be obvious on comparing the expression above with the sieve in the example above. For clarity we represent all innermost pairs of parentheses () by a blob •, and miss out the outermost pair and any other pair which is uniquely determined by the condition that the depth should be d. For example we would abbreviate the example above to (• • •)(•(••)).

The third way of representing sieves of width n and depth d is as rooted trees such that all branches are of length d, there are n "leaves", and all the edges going upwards from any given vertex are totally ordered. There is one vertex of the tree for every pair of parentheses (or to every equivalence class of every equivalence relation), and an edge between vertices if one of the vertices corresponds to a pair of parentheses which is a maximal pair inside the parentheses corresponding to the other vertex.

The number of sieves of width $n > 0$ and depth d is d^{n-1}, as it is easy to see by induction on n.

We say that a sieve p is a refinement of q if every equivalence relation in q is also in p, and we make the sieves into a category by saying that there is a unique morphism from any sieve to any refinement. The category of sieves has an initial element, consisting of the sieve with just two equivalence relations E_0 and E_1.

The idea of a relaxed multilinear category is that instead of one space of multilinear maps from $A_1, \ldots, A_n$ to B we should have several spaces, one for each sieve of size n (so that in some sense the spaces of multilinear maps depend on "the order in which we calculate the product of elements of the A_i's"). The space corresponding to a sieve should be related to the space of any refinement of that sieve.

More precisely, a relaxed multi category is given by the following data:

(1) A set of objects.

(2) For each collection of objects $A_1, \ldots A_n, B$ ($n \geq 1$) we are given a functor from $Sieve_n$ to sets. The value of this functor at $p \in Sieve_n$ is denoted by $Multi_p(A_1, \ldots, A_n; B)$ and is called the set of multi morphisms from $A_1, \ldots, A_n$ to B of type p. (It can be thought of as some sort of space of Taylor series expansions of multilinear maps.) If p is the initial object of $Sieve_n$ then we call $Multi_p(A_1, \ldots; B)$ the set of multi morphisms and sometimes miss out p from the notation.

(3) If $A_{11}, \ldots, A_{n,m_n}, B_1, \ldots, B_n, C$ are objects, $p_i \in Sieve_{m_i}$ are sieves of the same depth,

and $q \in Sieve_n$, then we are given a composition map

$$\begin{aligned} &Multi_{p_1}(A_{11}, \dots, A_{1m_1}; B_1) \times \cdots \times Multi_{p_2}(A_{n1}, \dots, A_{nm_n}; B_n) \\ &\quad \times Multi_q(B_1, \dots, B_n; C) \\ \to &Multi_{q(p_1,\dots,p_n)}(A_{11}, \dots, A_{nm_n}; C) \end{aligned}$$

where $q(p_1, \dots, p_n)$ is the sieve given by first putting the sieves p_i "side by side" and then putting the sieve q "on top of them". In terms of parentheses, this means we replace the i'th innermost pair of parentheses of q by p_i, and in terms of trees we join the trees together by making the i'th leaf of q into the root of p_i.

These data should satisfy the following axioms, which we will state only vaguely as we do not use them later.

(1) Composition is associative.
(2) There is an identity morphism in $Multi(A; A)$ for all A, with the obvious properties.
(3) Composition is "compatible" with the morphisms in the categories $Sieve_n$.

If the spaces of multi morphisms are all abelian groups and composition is multilinear then we call the category a relaxed multilinear category.

If we are given isomorphisms from $Multi(A_1, \dots; B)$ to $Multi(A_{\sigma(1)}, \dots, B)$ for all σ in the symmetric group and these isomorphisms satisfy various obvious conditions we call the category a relaxed symmetric multilinear category. Similarly we can define relaxed braided multilinear categories.

Example 4.1. Any additive tensor category or multilinear category is a relaxed multilinear category, and in these cases the spaces of multilinear maps do not depend on the choice of sieve p. (In other words the functors from $Sieve_n$ take all morphisms of $Sieve_n$ to the identity morphism.)

We can define associative rings in any relaxed multilinear category as follows. An associative ring consists of an object A and maps $f_n \in Multi(A, A, \dots A; A)$, $f_1 =$ identity, for all $n \geq 1$ (with n copies of A mapping to A) such that any composition of these maps f_m is the image of some f_n under the map from multilinear maps to compositions of multilinear maps. Informally we can think of the associativity property in a relaxed multilinear category like this: the products $(ab)c$ and $a(bc)$ cannot be compared directly because they lie in different spaces, but they can both be compared with abc using the map from the trivial sieve to the sieves generated by the equivalence relations $\{\{a, b\}\{c\}\}$ and $\{\{a\}\{b, c\}\}$.

Note that more complicated multilinear maps in a ring in a relaxed multilinear category are no longer always uniquely determined by compositions of bilinear maps, so we cannot sensibly define associativity of a bilinear map in $Multi(A, A; A)$ but can only define associativity of a sequence of multilinear maps f_n as above.

Similarly we can define commutative rings, Lie algebras, and so on in any relaxed symmetric multilinear category. As before, we need to specify all possible products of bilinear maps (as well as just bilinear maps) as part of the definitions of these things.

There are several variations of the definition of a relaxed multi category. For example, instead of using sieves we could use collections of intervals on $1, 2, \dots, n$ such that and two intervals in the collection are either disjoint or one contains the other, and all 1 and n point

sets are in the collection. (The union of all equivalence relations of a sieve is a collection with this property.) Then composition of multi maps is easier to axiomatize than if we use sieves, but it is harder to show that the representations of a vertex group satisfy the axioms.

The difference between an additive tensor category, a multilinear category, and a relaxed multilinear category can be illustrated by pretending that all multilinear maps are representable; we will of course denote the representing objects as tensor products. In a tensor category the object $A \otimes B \otimes C$ representing trilinear maps is isomorphic to $(A \otimes B) \otimes C$. In a multilinear category these need not be isomorphic even if both sides exist, but we would expect a canonical map from $A \otimes B \otimes C$ to $(A \otimes B) \otimes C$ corresponding to the composition of two bilinear maps being a trilinear map. In a relaxed multilinear category we would expect a map in the other direction from $(A \otimes B) \otimes C$ to $A \otimes B \otimes C$, corresponding to the fact that any trilinear map can be related to compositions of bilinear maps.

We can illustrate the differences by drawing the diagrams that have to commute for a bilinear map to be associative. For tensor categories, the following diagram has to commute:

$$\begin{array}{ccccc} (A \otimes A) \otimes A & & \cong & & A \otimes (A \otimes A) \\ \downarrow & & & & \downarrow \\ A \otimes A & \longrightarrow & A & \longleftarrow & A \otimes A \end{array}$$

For multilinear categories whose multilinear maps are representable the following diagram has to commute:

$$\begin{array}{ccccc} (A \otimes A) \otimes A & \longleftarrow & A \otimes A \otimes A & \longrightarrow & A \otimes (A \otimes A) \\ \downarrow & & & & \downarrow \\ A \otimes A & \longrightarrow & A & \longleftarrow & A \otimes A \end{array}$$

For relaxed multilinear categories whose multilinear maps are representable the following diagram has to commute. Notice that the arrows in the top row go in the opposite direction from the previous diagram. Moreover this is only the first of an infinite number of diagrams describing associativities of more copies of A that we need to define an associative product.

$$\begin{array}{ccccc} (A \otimes A) \otimes A & \longrightarrow & A \otimes A \otimes A & \longleftarrow & A \otimes (A \otimes A) \\ \downarrow & & \downarrow & & \downarrow \\ A \otimes A & \longrightarrow & A & \longleftarrow & A \otimes A \end{array}$$

A. Joyal pointed out to me the following similarities between Stasheff's A_∞ spaces and associative objects in a relaxed multilinear category. An A_∞ space is a sort of space with a product that is associative up to homotopy, and the homotopies are well defined up to higher homotopies, and so on. More precisely an A_∞ space A is a space A together with maps $K_n \times A^n \mapsto A$ for all $n \geq 2$, where K_n is a certain $n-2$-dimensional cell complex whose cells correspond to topological types of rooted trees with n leaves together with an ordering on the branches going upwards from each node. These are almost the same as sieves as defined above; the only difference is that in a sieve all the branches have the same height, and this seems to be a minor technical requirement which could probably

be removed by some slight changes of definition. Moreover the boundary maps in the complexes K_n correspond to the maps between sieves given by refinement. By adjointness we can think of an A_∞ space as a space A with maps from A^n to A^{K_n} for all n (satisfying various conditions) which is very similar to the definition of an associative object in the relaxed multilinear category of representations of a vertex group G, except that we use A^{K_n} instead of the spaces of singular A valued functions on G^n parameterized by sieves.

5. The representations of a vertex group.

In this section we will describe how to make the representations of a vertex group G into a symmetric relaxed multilinear category, so that it is possible to define commutative ring objects in this category. We let G be a vertex group with group ring H and ring of singular functions K.

The category C of representations of G is defined to be the category of representations of the group ring H. (Of course the multilinear maps in these two categories will usually be different!) In particular the category C has tensor products and Hom functors, as it is just the representations of some cocommutative Hopf algebra. There is a functor Γ from C to R-modules taking any G module to its fixed point set. Note that $\mathrm{Hom}_R(x,y)$ is an object of C, while $\mathrm{Hom}_G(x,y)$ is the morphisms from x to y which is just an R module. The relation between them is $\mathrm{Hom}_G(x,y) = \Gamma(\mathrm{Hom}_R(x,y))$.

We will first define the multilinear maps from $A_1, \dots, A_n$ to B in C. To motivate this we first describe the H-multilinear maps in a slightly unusual way. The R-multilinear maps are just the elements of $\mathrm{Hom}_R(A_1 \otimes \cdots \otimes A_n, B)$, and the H multilinear ones are just the H invariant elements of this space. Now consider the map taking $a_1 \in A_1, \dots, a_n \in A_n$ to $a_1^{g_1} a_2^{g_2} \cdots a_n^{g_n} \in B$ for $g_i \in H$. For fixed a_i this can be thought of as a G invariant B valued function on G^n, by which we mean an element of $\mathrm{Hom}_H(H \otimes \cdots \otimes H, B)$. Moreover this map from $A_1 \otimes \cdots \otimes A_n$ to B valued functions on G^n is G^n-invariant. So to summarize, the multilinear maps can be thought of as G^n invariant maps from $A_1 \otimes \cdots \otimes A_n$ to the G-invariant B-valued functions on G^n.

We will define G multilinear maps in the same way, except that we allow the functions on G^n to have "singularities". Informally we want this space to be the space of functions on G^n which are allowed to have singularities "of type K" whenever two of the components of G^n are equal. The formal definition goes as follows. The space $\mathrm{Hom}(H^n, B)$ of functions from G^n to B is a module over the ring H^{n*} of functions on G^n. For each $1 \le i < j \le n$ there is a homomorphism from H^* to H^{n*} induced by the map from H^n to H taking $g_1 \otimes \cdots \otimes g_n$ to $g_i S(g_j)$ (= "$g_i g_j^{-1}$"). We define the localization of a module M over H^{n*} at (i,j) to be $M \otimes_{H^*} K$, where M is made into an H^* module using the homomorphism f_{ij} from H^* to H^{n*}.

Definition 5.1. *We define the space $Fun(G^n, B)$ of singular functions from G^n to B to be the localization at all $(i,j), 1 \le i < j \le n$ of the space of functions from G^n to B.*

The module $\mathrm{Hom}(H^n, B)$ also appears in the computation of the cohomology of the G-module B using the homogeneous standard resolution: the cohomology groups of B are just the cohomology groups of a complex whose terms are the G-invariant elements $\mathrm{Hom}(H^n, B)^G$. We will not use this connection in this paper.

Example 5.2. Suppose G is the vertex formal group of example 3.4, so that $H^* = R[[x]]$ and $K = R[[x]][x^{-1}]$. Then the functions from G^3 to B can be identified with the space of formal power series $B[[x_1, x_2, x_3]]$ in 3 variables with values in B. The localization of this at (i, j) for $1 \leq i < j \leq 3$ is just $B[[x_1, x_2, x_3]][(x_i - x_j)^{-1}]$, so the space of singular functions from G^3 to B is the space

$$B[[x_1, x_2, x_3]][(x_1 - x_2)^{-1}, (x_2 - x_3)^{-1}, (x_1 - x_3)^{-1}].$$

Definition 5.3. *The G-module $Multi_R^K(A_1, \ldots, A_n, B)$ of multilinear maps is defined to be the space of H^{n*} invariant maps from $A_1 \otimes \cdots \otimes A_n$ to the space of singular B-valued functions on G^n. The R-module $Multi_G^K(A_1, \ldots, A_n, B)$ of G invariant singular multilinear maps is defined to be the module of G-invariant elements of $Multi_R^K(A_1, \ldots, A_n, B)$.*

We have defined the 4 different spaces of multilinear maps $Multi_R(A_1, \ldots, A_n, B)$, $Multi_G(A_1, \ldots, A_n, B)$, $Multi_R^K(A_1, \ldots, A_n, B)$, and $Multi_G^K(A_1, \ldots, A_n, B)$. The first two are the "usual" spaces of multilinear and invariant multilinear maps over the Hopf algebra H, and the other two are the corresponding spaces of singular multilinear maps.

To make C into a relaxed multilinear category we not only have to define the multilinear maps, but also the spaces $Multi_p(A_1, \ldots, B)$ for nontrivial sieves p, and the maps between them for $p < q$, and the compositions of multilinear maps.

We first define the multilinear maps of type p, for p in the sieve category $Sieve_n$. These can be thought of informally as various sorts of Taylor series expansions of multilinear maps.

Definition 5.4. *The space $Fun_p(G^n, B)$ of singular B-valued functions of type p is defined recursively as follows. Represent p as a tree of depth d, consisting of n leaves attached to a tree q of depth $d-1$. Suppose that X is the space of singular B-valued functions of type q. Then $Fun_p(G^n, B)$ is obtained from X in two steps: first we construct the space Y of all G^n valued functions with values in X, then we localize Y at all pairs i, j for i and j leaves of p joined to the same leaf of q.*

This space can be thought of as some space of singular functions "of type p" on G^n with values in B, where n is the number of nodes of the tree p. This space is also acted on by G^n.

Example 5.5. With G as in example 3.4, we have

$$\begin{aligned}
Fun_{(\bullet\bullet\bullet)}(G^3, B) &= B[[x_1, x_2, x_3]][(x_1 - x_2)^{-1}, (x_2 - x_3)^{-1}, (x_1 - x_3)^{-1}] \\
Fun_{((\bullet\bullet)\bullet)}(G^3, B) &= B[[x_1, x_2]][(x_1 - x_2)^{-1}][[x_3]][x_3^{-1}] \\
Fun_{(\bullet(\bullet\bullet))}(G^3, B) &= B[[x_2, x_3]][(x_2 - x_3)^{-1}][[x_1]][x_1^{-1}]
\end{aligned}$$

Definition 5.6. *The G-module $Multi_p^K(A_1, \ldots, A_n, B)$ of singular multilinear maps of type p is defined to be*

$$\mathrm{Hom}_{H^n}(A_1 \otimes \cdots \otimes A_n, Fun_p(G^n, B)).$$

The R-module $Multi^K_{Gp}(A_1, \dots, A_n, B)$ of G-invariant singular multilinear maps of type p is defined to be the G-invariant elements of $Multi^K_p(A_1, \dots, A_n, B)$.

The composition of multilinear maps of types $p_1, \dots, p_n$ and q is easily checked to be a multilinear map of type $q(p_1, \dots, p_n)$.

Finally we have to describe the map from multilinear maps of type p to multilinear maps of type q whenever $p < q$. The only possible problem in defining this map is showing that the localizations extend. The way to do this is best explained by an example. If we have an expression like $(x_1 - x_2)^{-k}$ we can expand it as a series

$$\sum_{i,j \geq 0} C(i,j)(x_1 - y_1)^i (y_1 - y_2)^{-k-i-j} (y_2 - x_2)^j$$

for some constants $C(i,j) = (-k)!/i!j!(-k-i-j)!$. In general there is an analogue of this power series expansion for any cocommutative Hopf algebra H.

6. G vertex algebras.

In this section we give the definition of a G vertex algebra.

We start with some motivation for the definition. Suppose that V is an associative algebra acted on by a group G, with the action of $g \in G$ on $v \in V$ written as v^g. For each fixed $v_1, \dots, v_n \in V$ we can form the products $v_1^{g_1} v_2^{g_2} \cdots v_n^{g_n}$, which we think of as a function from G^n to V. These function have the following properties:

1. Identity: $\cdots v_n^{g_n} 1^g v_{n+1}^{g_{n+1}} \cdots = \cdots v_n^{g_n} v_{n+1}^{g_{n+1}} \cdots$
2. Associativity and G invariance:

$$(v_{11}^{g_{11}} v_{12}^{g_{12}} \cdots)^{g_1} (v_{21}^{g_{21}} v_{22}^{g_{22}} \cdots)^{g_2} \cdots = v_{11}^{g_{11} g_1} v_{12}^{g_{12} g_1} \cdots v_{21}^{g_{21} g_2} v_{22}^{g_{22} g_2} \cdots$$

Conversely if we have a module V over a ring R together with functions from G^n to V as above which are multilinear in the v_i's then we can define the structure of an associative G algebra on V by defining the product $v_1 v_2$ to be $v_1^1 v_2^1$ and defining the G action on V by v^g. It is easy to check that this means that modules V with a set of functions as above are equivalent to associative algebras with an action of G.

We now define a G vertex algebra where G is a vertex group. We defined a vertex group earlier as a cocommutative Hopf algebra over R with a vertex structure, but we can think of it informally as being a group together with some ring of "singular" functions on it. A G vertex algebra is informally defined to be a module over R together with multilinear singular maps denoted by $v_1^{g_1} v_2^{g_2} \cdots v_n^{g_n}$ from G^n to V for all $v_1, \dots, v_n \in V$, such that the axioms above are satisfied. In other words we just copy the strange definition of an associative G-ring in the paragraph above, and replace functions on G with singular functions on G. The formal definition, which makes precise what we mean by a singular function, is as follows.

Definition 6.1. *A G vertex algebra is an associative algebra in the relaxed multilinear category of G-modules.*

If we give G the trivial vertex structure such that the singular V-valued functions on G^n are defined to be the ordinary functions $\mathrm{Hom}(G^n, V)$ on G^n then we see from the

previous remarks that a G vertex algebra is exactly the same as an associative algebra acted on by G (although defined in a rather odd way).

It is now easy to extend many concepts in ring theory to G vertex algebras as in the following examples.

Example 6.2. A G invariant left ideal of a G vertex algebra V is an R submodule I such that $v_1^{g_1} v_2^{g_2} \dots v_n^{g_n}$ is a singular map from G^n to I, whenever $v_1, \dots, v_{n-1} \in V$, $v_n \in I$. Similarly we can define right and two sided ideals.

Example 6.3. A left module over the G vertex algebra V is an R module M together with multilinear singular maps $v_1^{g_1} \cdots v_n^{g_n} m$ from G^n to M for all $v_i \in V, m \in M$, satisfying the obvious axioms about associativity and the identity element. This is the definition of a module without an action of G; we can define modules with an invariant action of G in the same way by using multilinear singular maps $v_1^{g_1} \cdots v_n^{g_n} m^g$ from G^{n+1} to M instead. We can define submodules, homomorphisms, and multilinear maps of modules in the obvious way. Unlike the case of rings there seems to be no reason in general why multilinear maps should be representable (though they often are in practice), but when they are representable we can use this to define tensor products of modules. (The problem of when this tensor product is associative seems rather hard.)

Example 6.4. We can define a commutative G vertex algebra to be one such that the multilinear singular functions are invariant under the obvious action of the symmetric group. We can define supercommutative G vertex algebras in a similar way.

Example 6.5. If V and W are two G vertex algebras then their tensor product is also a G vertex algebra, with the G vertex algebra structure defined by

$$(v_1 \otimes w_1)^{g_1} \cdots = (v_1^{g_1} \cdots) \otimes (w_1^{g_1} \cdots).$$

If V and W are commutative then so is $V \otimes W$. (Note that if V and W are supercommutative then there should be some extra signs $(-1)^{\deg(v_i)\deg(w_j)}$ for $i < j$ added to the definition of the product in $V \otimes W$.) More generally, if V and W are G_1 and G_2 vertex algebras, then $V \otimes W$ is a $G_1 \otimes G_2$ vertex algebra. If $G_1 = G_2 = G$ then there is a homomorphism from G to $G \otimes G$, which can be used to turn $V \otimes W$ into a G vertex algebra as in the first part of this example.

Example 6.6. If G is the vertex group of example 3.4 then it is easy to check that commutative G vertex algebras are the same as vertex algebras (as defined in [K, 1.3] for example). In fact if V is a vertex algebra and $v_1, \dots, v_n$ are any elements of it, and $v_i(z)$ is the vertex operator of v_i, then $v_1(z_1) \cdots v_n(z_n)1$ is an element of

$$V[[z_1, \dots, z_n]][\prod_{i<j} (z_i - z_j)^{-1}]$$

and hence defines a singular multilinear map from $V \times \cdots \times V$ to V, and it is not hard to check that the axioms for a vertex algebra in [K, 1.3] are equivalent to saying that this sequence of multilinear maps makes V into a commutative G vertex algebra. The field algebras of [K, 4.11], which are "non-commutative" vertex algebras, seem to be the same as associative G vertex algebras for G as in example 3.4.

Example 6.7. If G is the vertex group whose underlying Hopf algebra is the universal enveloping algebra of the Virasoro algebra, with the ring of singular functions given by

inverting an element as in example 3.4, then G vertex algebras are more or less the same as vertex operator algebras, provided that we restrict to the positive energy representations of the Virasoro algebra.

Example 6.8. If G is the product of two copies of the vertex group of example 3.4, then G can be thought of as a vertex group corresponding to 2 dimensional quantum field theories. A typical G vertex algebra can be constructed by taking the tensor product of 2 vertex algebras; this corresponds to the usual decomposition of fields in 2 dimensional field theory into their left moving and right moving parts. We can do the same thing if G is the product of 2 copies of the vertex group in example 6.7; this gives the vertex group underlying much of conformal field theory and string theory, and as before we can construct many examples by taking a tensor product of 2 vertex operator algebras.

One reason why nontrivial examples of G vertex algebras are hard to find is that there are almost no examples of finite dimensional G vertex algebras, except for those that are really just associative algebras. It is not hard to see why this should be so: a nontrivial G vertex algebra should have some sort of singularity, say a pole of order 1, otherwise it would just be an associative algebra. However from a pole of order 1 we can usually construct poles of all orders by algebraic operations, so we get an infinite dimensional space of possible singularities. This is not possible unless the G vertex algebra we started with is infinite dimensional.

7. Vertex operators.

The most powerful and general way of constructing G vertex algebras is to write down a set of vertex operators satisfying the "normal ordering" condition below. This is closely related to the well known method of constructing vertex algebras by writing down a set of commuting vertex operators; see [K, 4.5]. For the rest of this section we suppose that G is a vertex group with group ring H. We now define several rings of "singular functions" and some modules over them.

If V is an R module we define $V[[x]]$ to be $Hom_R(H, V)$ (where x stands for $\dim(G)$ variables). If G is a formal group this is isomorphic as an R module to the space of power series in $\dim(G)$ variables with coefficients in V, and is an H module (using the trivial action of H on V). Similarly we define $V[[x_1, \dots, x_n]]$ to be $Hom_R(H \otimes \cdots \otimes H, V)$, which is isomorphic to the space of power series in $n \dim(G)$ variables with coefficients in V and is an $H \otimes \cdots \otimes H$ module. It is also a module over the ring $R[[x_1, \dots x_n]] = \mathrm{Hom}(H \otimes \cdots \otimes H, R)$. We define $V((x_1, \dots, x_n))$ to be $V[[x_1, \dots, x_n]]$ localized at all $1 \le i \le n$ and at all pairs $1 \le i < j \le n$ as in section 5. For example, if G is as in example 3.4 then

$$V((x_1, x_2)) = V[[x_1, x_2]][x_1^{-1} x_2^{-1} (x_1 - x_2)^{-1}].$$

Definition 7.1. *A vertex operator on an R module A is an R linear map from A to $B((x)) = Hom_R(H, B) \otimes_{H^*} K$. More generally a vertex operator from $A_1 \times \cdots \times A_n$ to B is an R multilinear map from $A_1 \times \cdots \times A_n$ to $B((x_1, \dots, x_n))$.*

Example 7.2. With G as in example 3.4, a vertex operator from A to B is just a linear map from A to the space of formal Laurent series in B.

A vertex operator on A can be thought of informally as a singular function from G to operators from A to B.

It is not usually true that the composition of vertex operators is a vertex operator.

Definition 7.3. *We say that a sequence of vertex operators $v_1, \dots, v_n$ is* **compatible** *if their composition is the image of a unique singular multilinear map. More precisely suppose each v_i is a vertex operator from A_{i-1} to A_i, in other words an R linear map from A_{i-1} to $A_i((x_i))$. Then the composition is automatically a linear map from A_0 to $A_n((x_n)) \cdots ((x_1))$, and the condition that the composition is the image of a unique singular multilinear map means that it is in the image of a unique map from A_0 to $A_n((x_n, \dots, x_1))$. We say that a set S of vertex operators on A is compatible if any finite sequence of vertex operators in S is compatible.*

Definition 7.4. *We say that a vertex operator from A to B is a* **creation vertex operator** *if it is (induced by) a linear map from A to $B[[x]] = Hom(H, B)$.*

Definition 7.5. *We say that a vertex operator from A to B is an* **annihilation vertex operator** *if it factors through $B \otimes_R K$ via the natural map from $B \otimes_R K = B \otimes_R H^* \otimes_{H^*} K$ to $\mathrm{Hom}(H, B) \otimes_{H^*} K$.*

Roughly speaking, a creation vertex operator is a vertex operator that is nonsingular, and an annihilation vertex operator is one that is "rational" rather than "transcendental".

Definition 7.6. *We say that a product of vertex operators is* **normally ordered** *if for some $k \geq 0$ the first k operators in the product are creation vertex operators and the remainder are annihilation vertex operators.*

Definition 7.7. *We say that a set of vertex operators S on an R-module A satisfies the* **normal ordering condition** *if the composition of any two vertex operators in S can be written as a linear combination of normally ordered products of pairs of vertex operators in S.*

Lemma 7.8. *If a set of vertex operators satisfies the normal ordering condition then it is a compatible set of vertex operators.*

Proof. The main point is that any normally ordered product of annihilation and creation operators is a vertex operator. It follows immediately that if a set of vertex operators satisfies the normal ordering condition then any composition of vertex operators in S can be written as a linear combination of normally ordered compositions of vertex operators in S, by repeatedly rewriting product of operators so that creation vertex operators occur on the left, and hence is a vertex operator. So the set S is a compatible set of vertex operators. This proves lemma 7.8.

The point of the preceding definitions is the following theorem, which will be the main tool for finding compatible sets of vertex operators. It is an analogue of the trivial theorem that any set of operators on a module generates an algebra acting on the module.

Theorem 7.9. *Any compatible set of vertex operators on a module A generates a G vertex algebra acting on A.*

If A is a module over a commutative ring V which is generated as a V module by an element 1_A, then A obviously has the structure of a commutative ring, isomorphic to the kernel of V by the ideal of elements annihilating 1_A. The next theorem is a generalization of this to vertex algebras.

Theorem 7.10. *Suppose an R module A is acted on by a commutative G vertex algebra V, and suppose that A contains an element 1_A fixed by G and such that A is generated as a V-module by 1_A (which means that there is no proper G-submodule of A containing 1_A and acted on by V) and such that the vertex operator of any element of V applied to 1_A is nonsingular. Then the module A has a unique structure of a commutative vertex algebra such that the identity is 1_A and such that the action of any element of V on A is given by the action of some element of A.*

Proof. We define a map from V to A by mapping any v to $v^1(1_A)$ (which is well defined as we assumed $v^g(1_A)$ to be nonsingular in g). This is a homomorphism of V modules, and as V is commutative the kernel is a two sided ideal of V, so the image of V is a commutative vertex algebra. As this image is a V submodule containing 1_A, it must be the whole of A by assumption. Therefore A has the structure of a commutative G vertex algebra. This proves theorem 7.10.

The usual method of using the theorems in this section to construct G vertex algebras goes as follows. First we write down a lot of annihilation and creation operators satisfying the normal ordering condition. Then we try to find a commuting set of vertex operators in the vertex algebra they generate. There is often then an obvious element 1_A satisfying the conditions of theorem 7.10, so this makes A into a commutative G vertex algebra.

Example 7.11. We show how to construct the vertex algebra of a lattice using theorem 7.10. (See [K, 5.4] for a similar construction.) We assume for simplicity that all inner products in the lattice L are even. We let G be the vertex group of example 3.4 with underlying Hopf algebra H the ring spanned by the elements $D^{(i)}$. We let V be the universal commutative H-algebra generated by the group ring $R[L]$ of the lattice L. For each α in L we define a creation operator $e^{\alpha+}(z)$ to be multiplication by the element $\sum z^i D^{(i)}(e^\alpha)$. We define the annihilation operator $e^{\alpha-}$ to be the homomorphism of rings with derivation from V to $V[z,z^{-1}]$ taking e^β to $z^{(\alpha,\beta)}e^\beta$. These annihilation and creation operators generate a (non commutative) G vertex algebra acting on V, by theorem 7.9. The operators $e^\alpha(z) = e^{\alpha+}(z)e^{\alpha-}(z)$ all commute with each other, and applying theorem 7.10 shows that they induce the structure of a vertex algebra on V. This is the usual vertex algebra of a lattice L.

8. Free field G vertex algebras.

In this section we construct some examples of G vertex algebras closely related to generalized free quantum fields.

To keep notation simple we will first describe a special case of the construction with just one (scalar) field. Afterwards we will list various ways to generalize it. We assume that the underlying Hopf algebra of G is the universal enveloping algebra of the Lie algebra of the group of translations of spacetime, so it is a polynomial algebra generated by the elements $\frac{\partial}{\partial x_i}$ for $1 \le i \le d$. The dual H^* is the ring of formal power series $R[[x_1, \dots, x_d]]$. We will take K to be $H^*[(-x_1^2 + \cdots + x_n^2)^{-1}]$ (so we allow functions to have singularities along the light cone). We will say that an element f of H^* or K is even if $S(f) = f$ and odd if $S(f) = -f$, where S is the antipode with $S(x_i) = -x_i$. We fix an even element $\Delta(x)$ of K, which we call the propagator.

We let A be a free module over H generated by an element ϕ (which will be the

free field). The underlying space of the G vertex algebra V we are constructing is the symmetric algebra $S^*(A)$. This is an associative commutative algebra acted on by G, and we will call the product of elements a and b in this algebra the normal ordered product of a and b and denote it by $:ab:$

We will construct an annihilation vertex operator $\phi^-(x)$ and a creation vertex operator $\phi^+(x)$, and we define $\phi(x)$ to be $\phi^-(x)+\phi^+(x)$. We define the annihilation operator $\phi^-(x)$ to be the derivation of $R[[x]]$ algebras from $V[[x]]$ to $V[[x]]$ which commutes with the action of G and such that

$$\phi^-(x)(\phi) = \Delta(x).$$

We define the creation operator from V to $V[[x]]$ by

$$\phi^+(x)(v) = \sum_i x^i D^{(i)}(\phi)v.$$

Lemma 8.1. *The annihilation and creation operators satisfy the following commutation relations.*

$$\begin{aligned}[\phi^+(x), \phi^+(y)] &= 0\\ [\phi^-(x), \phi^+(y)] &= \Delta(x-y)\\ [\phi^-(x), \phi^-(y)] &= 0\\ [\phi(x), \phi(y)] &= 0\end{aligned}$$

Proof. The first equality is trivial because the ring V is commutative, and the fourth follows from the first three and the fact that Δ is even. The third equality follows because $[\phi^-(x), \phi^-(y)]$ is a commutator of 2 derivations on $V[[x,y]]$ and is therefore also a derivation, and it commutes with G and vanishes on V so it is zero. For the second equality we calculate that

$$\begin{aligned}\phi^-(x)(\sum_j y^j D^{(j)}(\phi)) &= \sum_j y^j D^{(j)}\Delta(x)\\ &= \Delta(x-y)\end{aligned}$$

which implies the second equality because $\phi^-(x)$ is a derivation. This proves lemma 8.1.

Theorem 8.2. *There is a unique structure of a commutative G vertex algebra on V such that the vertex operator $\phi(x)$ is the vertex operator of some element.*

Proof. This follows easily from lemma 8.1 and theorem 7.10 and some routine checks.

The elements of the G vertex algebra V are just the usual Wick polynomials in ϕ, and the vertex algebra product of elements of V is just the usual expansion of products of Wick polynomials in terms of normal ordered products. See [S-W], section 3.2, for examples.

The construction above can easily be generalized in several ways as follows.

1. We can allow more than one field ϕ, in which case Δ should be changed to a function $\Delta_{\phi\psi}$ depending on the fields ϕ and ψ.
2. We can enlarge G to a semidirect product of spacetime translations with some other group, such as the Lorentz group. This other group will act on the space of fields in 1 above, so we can define spinor fields, vector fields, and so on.

3. If Δ satisfies some differential equation (such as the wave equation), in other words if it is annihilated by some element D of H, then $D(\phi)$ generates a proper ideal of the G vertex algebra V so we can quotient out by it.
4. Suppose that we have a function $\Delta_n(g_1, g_2, \dots, g_n)$ of n variables $g_i \in G$ which are g invariant, with Δ_n nonsingular for $n > 2$. Then we can define a new G vertex algebra structure on $S^*(A)$ so that $\phi(x_1)\phi(x_2)\cdots\phi(x_n)$ to be $:\phi(x_1)\phi(x_2)\cdots\phi(x_n):$ $+\Delta(x_1, \cdots, x_n)$ More generally we define products of the form $\phi^{n_1}(x_1)\phi^{n_2}(x_2)\cdots$ to be a sum of terms, each of which is formed by repeatedly pulling out n factors of the form $\phi(x_i)$ with not all i's the same, and replacing them with Δ_n. Derivatives of ϕ are handled by differentiating Δ. More generally we can assume that we are given functions Δ_n for all $n \geq 2$ as above. The functions Δ_n are the "irreducible n-point functions".
5. In example 4 we need to assume that the functions Δ_n are nonsingular for $n > 2$ in order to fit into the framework of G vertex algebras. We can allow the functions Δ_n to be singular (but we have to assume that if we identify some but not all of the variables of Δ with each other then the result is a well defined function) if we are willing to enlarge the notion of a vertex group. In particular we have to allow singular multilinear functions to have more general sorts of singularities, so the more general definition of a vertex group should specify not just the singular functions of one variable, but also singular functions of many variables. We will leave the precise definition as an exercise for the reader.

Example 5 above gives examples corresponding to nontrivial quantum field theories defined perturbatively, which corresponds to the fact that the G vertex algebras are only defined over the ring of formal power series in the coupling constants.

9. The main identity.

In this section we prove an identity for commutative G vertex algebras where G is a vertex formal group associated to some algebraic groups over $\mathbf{C}$. Roughly speaking, the identity says that if we have a vertex differential operator acting on a commutative vertex algebra, then integrating it over a cycle of dimension n increases its order as a differential operator by at most n.

In the case of classical vertex algebras (when G as in example 3.4 is associated to the one dimensional additive algebraic group) this identity we will prove is equivalent to the identities originally used to define vertex algebras [B] (see [K]). We will start off by explaining why, as motivation for the proof. The vertex algebra identity states that

$$\sum_{i\in\mathbf{Z}} \binom{m}{i} (u_{q+i}v)_{m+n-i}w = \sum_{i\in\mathbf{Z}} (-1)^i \binom{q}{i} (u_{m+q-i}(v_{n+i}w) - (-1)^q v_{n+q-i}(u_{m+i}w)).$$

For simplicity we will only discuss the case $m = n = q = 0$, when it becomes (one version of) the Jacobi identity $(u_0v)_0w = u_0(v_0w) - v_0(u_0w)$; the general case follows from a similar argument after including a suitable rational function of x, y and z. This Jacobi identity can be deduced from the fact that u_0 is a differential operator of degree 1 as follows. The fact that u_0 is a differential operator of degree 1 means that $(u_0)v(y) - v(y)u_0$ is a differential operator of degree 0 and must therefore be of the form $t(y)$ for some $t \in V$;

putting $y = 0$ then shows that $t = u_0 v$, so that $u_0(v(y)w) - v(y)u_0 w = (u_0 v)(y)w$, and now integrating y around 0 gives the Jacobi identity above. This justifies the statement that the main identity for a vertex algebra is equivalent to the fact that integrating $u(x)$ along a 1-cycle is a vertex differential operator of degree 1.

To prove that the integral u_0 of $u(x)$ along a 1-cycle is a vertex differential operator of degree 1 we have to show that the double commutator

$$[[u_0, v(y)]w(z)]$$

is 0, and each term in the double commutator can be written as the integral of $u(x)v(y)w(z)$ along a suitable cycle in the x plane with the points 0, y, and z removed. More precisely, if $C_{a,b,c\ldots}$ is a 1-cycle going once clockwise around the points $a, b, c \ldots$ and not containing other elements of the points $0, x, y$ then

$$C_{0,y,z} - C_{0,y} - C_{0,x} + C_0 = 0$$

in the first homology group, which gives us the relation

$$(u_0)v(y)w(z) - w(z)(u_0)v(y) - v(y)(u_0)w(z) + v(y)w(z)u_0 = 0.$$

In other words, the fact that u_0 is a vertex differential operator follows from linear relations between elements of a certain homology group.

In the higher dimensional case we proceed in a similar way, showing that certain operators are vertex differential operators of higher degree using relations between elements of homology groups, and this can be thought of as a higher dimensional generalization of the vertex algebra identity. At first sight there seems to be a serious problem in carrying out this program: the higher dimensional spaces whose homology groups we work with are not only rather complicated, but there are far too many of them to calculate all their homology groups. Rather surprisingly, we do not need to know the exact homology groups, and we do not even need to find any explicit relations. It turns out that the identities we want to prove follow just from the existence of "sufficiently many" relations. We will prove the existence of enough relations by bounding the dimensions of the homology group and using the fact that if we have a set of elements in the homology group with cardinality greater than its rank then there must be a nontrivial linear relation between them.

We let G be a finite dimensional real vertex formal group. The underlying formal group of G is the formal group of some connected real algebraic group, and we assume that the vertex structure on G is given by inverting some element of the coordinate ring of this algebraic group (so that K is of the form $RG^*[1/p] = R[[x_1, \ldots]][1/p]$ for some polynomial p in $x_1, \ldots$). We let U stand for a unipotent algebraic subgroup of G of some dimension n which is not contained in the divisor of p. As U is unipotent the exponential map from the Lie algebra of $U(\mathbf{C})$ to $U(\mathbf{C})$ is a diffeomorphism, so that $U(\mathbf{C})$ is a vector space over $\mathbf{C}$ with some group structure (which is of course different from the additive group structure if U is not abelian).

The function p restricts to a function on U with zero set given by some divisor D. We choose a compact n cycle $C \subset U(\mathbf{C})$ in the complement of the divisor D, which represents an element of $H_n(U(\mathbf{C}) - D)$ which we also denote by C.

We recall the definition of a differential operator on a ring. If S is a commutative algebra over a commutative ring R, then a differential operator D of order at most $n \in \mathbf{Z}$ is defined to be an operator which is zero if $n < 0$, and if $n \geq 0$ it is defined to be an operator such that $[D, s]$ is a differential operator of order at most $n-1$ for any element $s \in S$. A differential operator D is called normalized if $D(1) = 0$. The differential operators of order at most 0 are just multiplications by elements in the center of S, and normalized differential operators of order at most 1 are the same as derivations of S, or in other words operators such that $D(ab) = aD(b) + D(a)b$ for all $a, b \in S$. Every differential operator can be written uniquely as the sum of a normalized differential operator and multiplication by a constant. If D and E are differential operators of orders m and n then DE is a differential operator of order at most $m+n$, and $[D, E]$ is a differential operator of order at most $m+n-1$. In particular the normalized differential operators of order at most 1 form a Lie algebra acting on S, called the Lie algebra of derivations of S. A derivation of S can be thought of as an infinitesimal automorphism of S. We define vertex differential operators of commutative G vertex algebras in the same way.

Theorem 9.1. *If a is an element of a commutative G vertex algebra V with G as above, and n is the dimension of the subgroup U, then $\int_C a(z)d^n z$ is a vertex differential operator of order at most n. In other words if $a_0, \dots a_n \in V$ then*

$$[a_0(z_0), [a_1(z_1), [\dots, [a_n(z_n), \int_C a(z)d^n z] \cdots]]] = 0.$$

Note that $\int a(z)dz\, b(y) \neq b(y) \int a(z)dz$ even though $a(z)b(y) = b(y)a(z)$. The reason for this is that the two integrals are taken over different cycles in the subset where $a(z)b(y)$ is holomorphic, and these two cycles need not be homologous if $a(z)b(y)$ has singularities.

The proof of theorem 9.1 is in two steps. We first show that the number of linearly independent elements of the form

$$\left(\prod_{i \in S, 0 \leq i \leq N} a_i(z_i)\right) \int_C a(z)d^n z \left(\prod_{i \notin S, 0 \leq i \leq N} a_i(z_i)\right) \tag{9.2}$$

for the 2^{N+1} subsets S of $0, 1, 2, \dots, N$ is bounded by a polynomial in N of degree at most n (rather than the obvious bound 2^{N+1}) as N tends to infinity. This is because the number of linearly independent elements is bounded by the rank of certain homology groups, and we can bound these ranks by a polynomial in N. We then show that the possible relations between these elements are so restricted that this crude bound implies (and is even equivalent to) a single explicit relation, which is the one given in theorem 9.1.

Lemma 9.3. *Suppose $U(\mathbf{C})$ is a complex connected unipotent algebraic Lie group of dimension n and D is a closed algebraic subset. Then the rank of the homology group $H_n(U(\mathbf{C}) - g_1(D) \cup g_2(D) \cup \cdots \cup g_N(D))$ for elements g_i in general position is bounded by a polynomial of degree n in N (whose coefficients depend only on U and D).*

Proof. The one point compactification of $U(\mathbf{C})$ is a sphere of dimension $\dim(U(\mathbf{C})) = 2n$, so by Spanier-Whitehead duality it is sufficient to prove that the rank of the homology

group $H_{n-1}(g_1(D) \cup g_2(D) \cup \cdots \cup g_N(D))$ is bounded by a polynomial of degree n in N. There is a spectral sequence converging to the homology of $g_1(D) \cup g_2(D) \cup \cdots \cup g_N(D)$ whose E_2 term is given by the homology groups of all finite intersections of the $g_i(D)$'s (see [G, theorem 5.4.1 and the remarks in section 5.6]). The rank of $H_{n-1}(g_1(D) \cup g_2(D) \cup \cdots \cup g_N(D))$ is bounded by the sum of the ranks of the terms $H_{n-1-(k-1)}(g_1(D) \cap g_2(D) \cap \cdots \cap g_k(D))$ of the E_2 term of the spectral sequence of total degree n, and the rank of any homology group of any intersection of the $g_i(D)$'s is bounded by some constant, so it is sufficient to show that the number of ways of choosing k of the g_i's so that their intersection has vanishing $H_{n-1-(k-1)}$ is bounded by a polynomial in N of degree at most n. As the g_i's are in general position, the intersection of k of them has dimension at most $2(n-k)$, so that the H_{n-k}'th homology of this intersection vanishes whenever $n-k > 2n-2k$, or $k > n$. Hence the number of ways of choosing k of the g_i's so that their intersection has non vanishing $H_{n-1-(k-1)}$ is bounded by the number of subsets of N elements of size at most n, which is a polynomial of degree n in N. This proves lemma 9.3.

Lemma 9.4. *The number of linearly independent elements of the form 9.2 is bounded by a polynomial in N of degree n.*

Proof. Each term of the form 9.2 is given by integrating

$$a(z)a_0(z_0)\cdots a_n(z_n)d^n z$$

over some n-cycle in $G - z_0(D) \cup \cdots \cup z_N(D)$. By lemma 9.3 the rank of the homology group generated by these n-cycles is bounded by a polynomial in N of degree n, which proves lemma 9.4.

In particular there must be some nonzero relation for sufficiently large N, because the number of expressions of the form 9.2 increases like 2^N, but the number of linearly independent ones is bounded by some polynomial in N.

Lemma 9.5. *Suppose that the coefficients c_S are the coefficients of a nonzero relation of smallest possible degree N as above. Then $c_S = (-1)^{|S|}c_\emptyset$ for some nonzero constant $c_\emptyset$.*

Proof. We are given that

$$\sum_S c_S \left(\prod_{i \in S, 0 \le i \le N} a_i(z_i) \right) \int_C a(z) d^n z \left(\prod_{i \notin S, 0 \le i \le N} a_i(z_i) \right) = 0$$

for all $a_0, \ldots, a_N \in V$. But if we put $a_i = 1$, $z_i = 0$ we find a relation of smaller degree. All coefficients of this relation must be identically 0 as the c_S's were by assumption the coefficients of a nonzero relation of smallest degree. But the coefficients of this smaller relation are of the form $\pm(c_S + c_{S \cup i})$ for $S \subset \{1, \ldots, i-1, i+1, \ldots, N\}$, so $c_S = -c_{S \cup i}$. This implies that $c_S = (-1)^{|S|}c_\emptyset$ for all S. This proves lemma 9.5.

In particular this shows that $\int_C a(z)d^n z$ is a differential operator of some order. We now pin down the order more precisely by looking more carefully at the set of all possible relations.

Lemma 9.6. *Suppose that n' is the degree of the unique nonzero relation of smallest degree, as in lemma 9.5. Then the maximum number of linearly independent elements of the form 9.2 is exactly equal to the number of subsets of $\{0, 1, \ldots, N\}$ of size at most n' and is therefore a polynomial in N of degree exactly n'.*

Proof. We show that the set of elements of the form

$$\left(\prod_{i\in S, 0\le i\le N} a_i(z_i)\right) \int_C a(z) d^n z \left(\prod_{i\notin S, 0\le i\le N} a_i(z_i)\right)$$

with $|S| \le n'$ form a maximal linearly independent set of functions of the $a_i(z_i)$'s. In fact by using the relation of lemma 9.5 we see that any relation can be written as a sum of relations with at most n' factors to the left of the integral. To complete the proof of the lemma we have to show there are no nontrivial linear relations between these terms. Suppose there is a nontrivial relation with coefficients c_S, with $c_S = 0$ if $|S| > n'$. We can assume that we have chosen this relation so that the maximum value of $|S|$ with $c_S \neq 0$ is as small as possible. Now set $a_i = 1$, $z_i = 0$ for $i \notin S$. Then we get a relation of degree at most n' with a nonvanishing coefficient of $\left(\prod_{i\in S, 0\le i\le N} a_i(z_i)\right) \int_C a(z) d^n z$, which is impossible. Hence the maximal number of linearly independent elements is exactly equal to the number of subsets of size at most n' of a set of size N, which is a polynomial in N of degree exactly n'. This proves lemma 9.6.

We can now complete the proof of theorem 9.1. By lemma 9.4 the maximal number of linearly independent elements of the form 9.2 is bounded by a polynomial of degree at most n. In particular there must exist some relation of minimal degree n' by lemma 9.5. By lemma 9.6 we see that $n' \le n$, so by lemma 9.5 the relation 9.1 holds, in other words $\int_C a(z) d^n z$ is a vertex differential operator of order at most $n' \le n$. This proves theorem 9.1.

Theorem 9.1 can easily be generalized in several ways as follows.

1 The condition that C should lie inside some unipotent subgroup U can be removed; it is put in only because it simplifies the proof slightly and is satisfied in all the examples we use later.

2 The restriction that G is finite dimensional is also usually unnecessary; for example we could take G to be the formal group of the Virasoro algebra and take U to correspond to the 1 dimensional group generated by L_{-1}.

3 We can also look at vertex differential operators $a(x)$ and find that $\int_C a(z) d^n z$ is a vertex differential operator of order at most the order of a plus the dimension of U.

4 We can also look at vertex differential operators $a(x, y, z, \ldots)$ in several parameters, and find that $b(y, z, \ldots) = \int_C a(x, y, z, \ldots) d^n x$ is a vertex differential operator with order as above.

5 We can include a singular function f of several variables in the integrand without affecting the argument. For example if a is an element of a commutative G vertex algebra V with G as above, and n is the dimension of the subgroup U, then $\int_C a(z) f(z, z_0, \ldots, z_n) d^n z$ is a vertex differential operator of order at most n.

6 We do not need to restrict the vertex structure on G to be given by inverting a function p on an algebraic group, and we can allow things like the vertex structure generated by $f^{1/n}$ for some integer n.

7 Finally the restriction to vertex groups over the reals is unnecessary and theorem 9.1 can be generalized to vertex groups over any field R by using étale cohomology instead of singular cohomology.

10. G Vertex algebras and the Yang-Baxter equation.

In this section we show how to construct examples of G vertex algebras from solutions of the Yang-Baxter equation. The idea is to start with an G algebra, and deform it using a solution of the Yang-Baxter equation into a commutative G vertex algebra. (Note that in many examples we end up with a commutative G vertex algebra, even though the algebra we start with is not commutative!)

We first consider the case of associative algebras V. We say that a linear map R from $V \otimes V$ to $V \otimes V$ is an R matrix if it satisfies the following conditions:

1. $R_{12}R_{13}R_{23} = R_{23}R_{13}R_{12}$ (Yang-Baxter equation.) Both sides act on $V \otimes V \otimes V$, and R_{ij} means R acting on the i'th and j'th factors of the tensor product.
2. $R(1 \otimes v) = 1 \otimes v$, $R(v \otimes 1) = v \otimes 1$.
3. $R_{12}m_{12} = m_{12}R_{23}R_{13}$, $R_{12}m_{23} = m_{23}R_{12}R_{13}$, where m_{ij} is the product map from $V \otimes V \otimes \cdots V$ to $V \otimes \cdots V$ given by multiplying the i'th and j'th factors.

Note that we have added extra conditions saying that the R matrix is "compatible" with the algebra structure.

Lemma 10.1. *Suppose V is a commutative algebra and R is an R matrix for V. Then the bilinear map $m_{12}R_{12}$ is another associative algebra structure on V (with the same identity element 1).*

Proof. The only nontrivial thing to check is associativity, which follows from

$$\begin{aligned} & m_{12}R_{12}m_{23}R_{23} \\ =& m_{12}m_{23}R_{12}R_{13}R_{23} \\ =& m_{12}m_{12}R_{23}R_{13}R_{12} \\ =& m_{12}R_{12}m_{12}R_{12}. \end{aligned}$$

Now we assume that V is acted on by a group G. The group $G \times G$ then acts on $V \otimes V$ and on $Hom(V \otimes V, V \otimes V)$. We denote the image of the matrix R under $g \times g'$ by $R^{g,g'}$. We assume that R satisfies the following axioms. The first three are the obvious analogues of the ones above, the fourth is just the definition of the action of $G \times G$ on R, and the fifth is a sufficient condition for the G-invariance of the new algebra structure on V.

1. $R_{12}^{g_1,g_2}R_{13}^{g_1,g_3}R_{23}^{g_2,g_3} = R_{23}^{g_2,g_3}R_{13}^{g_1,g_3}R_{12}^{g_1,g_2}$ (Yang-Baxter equation.)
2. $R^{g_1,g_2}(1 \otimes v) = 1 \otimes v$, $R^{g_1,g_2}(v \otimes 1) = v \otimes 1$.
3. $R_{12}^{g_1,g_2}m_{12} = m_{12}R_{23}^{g_1,g_2}R_{13}^{g_1,g_2}$, $R_{12}^{g_1,g_2}m_{23} = m_{23}R_{12}^{g_1,g_2}R_{13}^{g_1,g_2}$.
4. $(R^{g_1,g_2}(u \otimes v))^{g_1',g_2'} = R^{g_1g_1',g_2g_2'}((u \otimes v)^{g_1',g_2'})$
5. $R^{gg_1,gg_2} = R^{g_1,g_2}$

If we want we can use the fifth axiom to define $R^g = R^{1,g}$ so that $R^{g_1,g_2} = R^{g_1^{-1}g_2}$, so R really only depends on one element of G.

As before, we can use R^{g_1,g_2} to twist a G invariant algebra product on V to get a new G-invariant algebra product.

We can now try to construct G vertex algebras in the same way, except that the matrix R^{g_1,g_2} is a singular function of g_1, g_2; more precisely, R^g should be an element of $Hom(V \otimes V, Hom(H, V \otimes V) \otimes_{H^*} K)$. The quantum groups literature contains many examples of solutions of equation 1 above with singularities for g_1, g_2 complex numbers, or more generally elements of some Riemann surface. These do not usually satisfy equation 3 because the space V is usually taken to be a finite dimensional vector space rather than an algebra, but equation 3 shows that there is at most one extension of R to any algebra generated by this finite dimensional space, such as the tensor algebra. This gives a large number of G vertex algebras constructed using solutions of the Yang-Baxter equation.

Example 10.2. We take G as in example 3.4, and show how to construct the vertex algebra of a lattice L using the method above. For simplicity we assume all inner products in L are even; the general case can be done in a similar way using a twisted group ring of L. We take V to be the universal commutative H algebra generated by the group ring $\mathbf{Z}[L]$ of L. There is a singular solution to the Yang-Baxter equation on $\mathbf{Z}[L]$, given by

$$R^{x,y}(e^a \otimes e^b) = (x-y)^{(a,b)} e^a \otimes e^b.$$

This extends uniquely to an R matrix on V satisfying the conditions above, which can be used to make V into a vertex algebra. This is the usual vertex algebra of an even lattice.

11. Cohomology of G vertex algebras.

We have seen earlier that many G vertex algebras can be constructed by deforming the product on a G-algebra. In this section we will show how to describe the infinitesimal deformations of a G-algebra into a G vertex algebra using G-equivariant cohomology of associative algebras.

We start by recalling how to classify the infinitesimal deformations of an associative algebra V over a commutative ring R. This means that we want to define an associative algebra structure on $V[\epsilon]$ over $R[\epsilon]$, where $\epsilon^2 = 0$. For simplicity we will work with associative algebras without identity elements; the main difference below if we use algebras with identity elements is that we should use normalized cochains instead of cochains. If we write the new algebra product as $ab + \epsilon f(a,b)$ then associativity is equivalent to

$$af(b,c) - f(ab,c) + f(a,bc) - f(a,b)c = 0$$

so f is just a 2-cycle for the standard complex used to calculate the associative algebra cohomology $H^2(V,V)$. (Recall that the n-cochains of the standard resolution for calculating the Hochschild cohomology groups $H^n(V,M)$ for a 2 sided module M over the associative algebra V are the multilinear maps from $A, A, \ldots, A$ to M; see [C-E].) The "trivial" deformations are those that are induced by an infinitesimal automorphism $a \mapsto a + \epsilon g(a)$ of the underlying R-module, when the corresponding 2-cocycle f is given by $f(a,b) = ag(b) - g(ab) + g(a)b$, in other words f is just the coboundary of the 1-cochain g.

Therefore the infinitesimal deformations of the algebra V are classified by the cohomology group $H^2(V,V)$ of the associative algebra V with coefficients in the 2-sided module V.

Now suppose that V is acted on by a group (or cocommutative Hopf algebra) G. Then we can calculate the infinitesimal deformations of the G-algebra V by using G-equivariant cohomology groups $H^2_G(V,V)$. The G-equivariant cohomology groups are calculated (and defined) by replacing the (standard) cochains above by the G-invariant cochains.

Finally suppose that G is a vertex group. We define the cohomology groups $H^n_G(V,M)$ by using the G-invariant singular multilinear functions from $V,V,\dots V$ to M in place of the multilinear functions from $V,V\dots,V$ to M. (The boundary operator is defined by the same formula as for associative algebras.) We then find by the argument above that the is a map from $H^2_G(V,V)$ for an associative G-algebra V to G vertex algebras over $R[\epsilon]$. The group $H^2_G(V,V)$ can be thought of a roughly the tangent space at A of the moduli space of G vertex algebra structure on A. Some examples of nontrivial elements of $H^2_G(V,V)$ can easily be obtained from the G vertex algebras of generalized free fields, because we can just take a 2-point function Δ of the form $\epsilon/(x-y)^2$ for example.

We can also define other sorts of cohomology groups involving vertex groups G, such as vertex group cohomology. We first note that most definitions in group cohomology (in particular the homogeneous and inhomogeneous standard complexes) can easily be extended to the case of arbitrary cocommutative Hopf algebras. We can now define the vertex group cohomology groups $H^n(G,M)$ for a vertex group G and a G-module M by replacing the spaces of multilinear maps $Hom(H,H\dots,H,M)$ in the homogeneous standard complex (where H is the underlying cocommutative Hopf algebra) by the corresponding spaces of singular multilinear maps.

12. Relation to quantum field theory.

There are several mathematical structures closely related to quantum field theory; for example, Wightman's axioms [S-W] and some closely related variations [G-J], Segal's axioms for a topological field theory [Se], and G vertex algebras as described above. We give an informal description of the relation between G vertex algebras and the other theories above. The four theories above are closely related but not exactly the same.

The relation with Wightman's axioms is easiest to describe. A quantum field theory satisfying Wightman's axioms is determined by its correlation functions [S-W], and if these correlation functions have "good" operator product expansions then they are the correlation functions of some G vertex algebra, with G the vertex group whose underlying Hopf algebra is the universal enveloping algebra of the Poincaré algebra, and where we allow some sort of singularities on the light cone. Unfortunately it is difficult to decide when a G vertex algebra comes from a quantum field theory satisfying the Wightman axioms. The main problem is in reconstructing the Hilbert space. It is usually not too hard to construct a real vector space with a symmetric bilinear form on it, but deciding when this form is positive (semi) definite is usually rather hard. (For example a special case of this is the problem of deciding which highest weight representations of the Virasoro algebra are unitary.)

Note that expressions like $\phi_1(x_1)\phi_2(x_2)$ are interpreted in quite different ways in quantum field theories and G vertex algebras: in quantum field theories we think of the ϕ_i's as distribution valued operators defined on a manifold containing points x_i, and the

product is a product of operators. In G vertex algebras the elements x_i are thought of as elements of some group and $\phi(x)$ should be thought of as the transform of the element ϕ under the action of the group element x. The product does not always make sense for fixed x_1 and x_2, but is only defined as some sort of singular function of x_1 and x_2.

The relation with quantum field theories that occur in physics is similar: provided that good operator product expansions exist, there is often a G vertex algebra with the same correlation functions. The main problem is that most realistic quantum field theories are only defined at the level of perturbation theory, in other words, the correlation functions are formal power series in the coupling constants that (probably) do not converge for any nonzero values. We can get round this by the following trick: we define the G vertex algebra over the ring of formal power series in the coupling constants. So many quantum field theories are now well defined mathematical objects: they are G vertex algebras over formal power series rings. Notice that this does not solve the important problem of making sense of these theories non perturbatively; all we have done is change the question of existence of quantum field theories into an equally hard question about properties of G vertex algebras over formal power series rings. More precisely, we would really like to construct some sort of moduli space of G vertex algebras, such that at points of its compactification the G vertex algebras somehow degenerate into the formal power series G vertex algebras above.

The relation of G vertex algebras with Segal's topological field theories is harder to describe; in fact, there seems to be no particularly easy way to go between them. The reason for this seems to be that G vertex algebras like to work with groups G and work best when spacetime can be regarded as a group (for example, if spacetime is flat). On the other hand, Segal's axioms work with arbitrary manifolds, most of which have nothing to do with groups. A good example is given by 2 dimensional conformal field theories. It is well known that vertex operator algebras (which are more or less vertex $Virasoro$-algebras) are good at describing the genus 0 part of a conformal field theory, but are bad at describing the higher genus part (although the genus 1 case has been pushed through by Zhu [Z]). This is because a genus 0 Riemann surface is more or less the additive group $\mathbf{C}$ (at least if a point is missed out), while Riemann surfaces of genus greater than 1 are not directly related to groups. The additive group $\mathbf{C}$ is of course more or less the "group" that acts on vertex algebras.

To summarize, quantum field theories satisfying the Wightman axioms, topological field theories, and G vertex algebras are related but different mathematical structures, each of which captures part but not all of the notion of a quantum field theory. Algebraic quantum field theories emphasize Lorentz invariance, locality, and unitarity, but have the disadvantage that it is extraordinarily difficult to construct nontrivial examples of them. G vertex algebras emphasize locality, Poincare invariance, and the operator product expansion, and can cope with the Feynman path integral (regarded as a trace on the G vertex algebra), but are not very good at dealing with unitarity or curved spacetimes. Segal's axioms emphasize the Feynman path integral and are particularly good at dealing with curved spacetimes.

References.

[A] E. Abe, Hopf algebras, Cambridge University Press, 1980.

[B-D] A. Beilinson, V Drinfeld, Chiral algebras I. 1996(?) preprint.

[B] R. E. Borcherds, Vertex algebras, Kac-Moody algebras, and the monster. Proc. Natl. Acad. Sci. USA. Vol. 83 (1986) 3068–3071.

[C-E] H. Cartan, S. Eilenberg, Homological algebra, Princeton U. P. 1956.

[G-J] J. Glimm, A. Jaffe, Quantum physics, Springer Verlag 1981.

[G] R. Godement, Théorie des faisceaux, Hermann 1964.

[K] V. G. Kac, Vertex algebras for beginners, University lecture series vol. 10, A.M.S. 1996.

[L-Z] B. Lian, G. Zuckerman, New perspectives on the BRST-algebraic structure in string theory, hep-th/9211072, Communications in Mathematical Physics 154, (1993) 613–64, and Moonshine cohomology, q-alg/950101, Finite groups and Vertex Operator Algebras, RIMS publication (1995) 87-11.

[Se] Segal, Graeme, Geometric aspects of quantum field theory. Proceedings of the International Congress of Mathematicians, Vol. I, II (Kyoto, 1990), 1387–1396, Math. Soc. Japan, Tokyo, 1991.

[S] J-P. Serre, "Lie algebras and Lie groups", 1964 lectures given at Harvard University. Second edition. Lecture Notes in Mathematics, 1500. Springer-Verlag, Berlin, 1992.

[So] Y. Soibelman, Meromorphic tensor categories, preprint q-alg/9709030.

[S-W] R. F. Streater, A. S. Wightman, PCT, Spin, statistics, and all that. Addison-Wesley 1964.

[Z] Zhu, Yongchang, Modular invariance of characters of vertex operator algebras. J. Amer. Math. Soc. 9 (1996), no. 1, 237–302.

arXiv:q-alg/9709033v1 22 Sep 1997

What is a vertex algebra?.

Arbeitstagung talk 1997
Richard E. Borcherds, *
D.P.M.M.S., 16 Mill Lane, Cambridge, CB2 1SB, England.
reb@dpmms.cam.ac.uk, http://www.dpmms.cam.ac.uk/~reb

The answer to the question in the title is that a vertex algebra is really a sort of commutative ring. I will try to explain this in the rest of the talk, and show how to use this to generalize the idea of a vertex algebra to higher dimensions. The picture to keep in mind is that a commutative ring should be thought of as somehow related to quantum field theories in 0 dimensions, and vertex algebras are related in the same way to 1 dimensional quantum field theories, and we want to find out what corresponds to higher dimensional field theories. This talk is an exposition of the paper q-alg/9706008, which contains (some of) the missing details. There is also some overlap with unpublished notes of Soibelman, which have just appeared on the q-alg preprint server as q-alg/9709030.

The relation of vertex algebras to commutative rings is obscured by the rather bad notation generally used for vertex algebras. Recall that for any element v of a vertex algebra V we have a vertex operator denoted by $V(v,z)$ taking V to the Laurent power series in V. I am going to change notation and write $V(v,z)u$ as $v^z u$. Let us see what several standard formulas look like in this new notation:

Old notation	New notation
$V(u,z)v$	$u^z v$
$V(V(a,x)b,y)c = V(a,x+y)V(b,y)c$	$(a^x b)^y c = a^{xy}(b^y c)$
$V(a,x)V(b,y)c = V(b,y)V(a,x)c$	$a^x b^y c = b^y a^x c$
$V(a,x)b = e^{xL_{-1}}(V(b,-x)a)$	$a^x b = (b^{x^{-1}} a)^x$
$V(1,x)b = b$	$1^x b = b$

The formulas on the right hand side are all easy to recognize: they are just standard formulas for a commutative ring acted on by a group G, where a, b, c are in the ring V, x, y, z are elements of the group G, and the action of $x \in G$ on $a \in V$ is denoted by a^x. This suggests that we should try to set things up so that vertex algebras are exactly the commutative rings objects over some sort of mysterious group-like thing G.

For simplicity we will work over a field of characteristic 0. This is not an important assumption; it just saves us from some minor technicalities about divided powers of derivations.

We will first look at the special case of vertex algebras such that all the vertex operators $V(a,x)$ are holomorphic. We show that such vertex algebras are the same as commutative algebras with a derivation D. The correspondence is given as follows. First suppose that V is a commutative algebra with derivation D. We define the vertex operator $V(a,x)$ by $V(a,x)b = \sum_{i\geq 0}(D^i a)bx^i/i!$. Conversely if V is a vertex algebra we define the product

* Supported by a Royal Society professorship.

by $ab = V(a,0)b$ and the derivation by Da = coefficient of x^1 in $V(a,x)b$. (We cannot really check that this turns commutative algebras into holomorphic vertex algebras and vice versa because we have not yet said exactly what the axioms for a vertex algebra are.)

In the new notation for vertex algebras above we would put $a^x = \sum_i x^i D^i a/i!$, $a^x b = \sum_i x^i D^i ab/i!$. Here we think of x as being an "element" of the one dimensional formal group $\hat{G}_a$. This formal group has as its formal group ring H the algebra of polynomials $k[D]$ and its coordinate ring is the ring of formal power series $k[[x]]$. (In characteristic 0 it does not matter whether we use Lie algebras or formal groups which are essentially equivalent, but in other characteristics formal groups are better than Lie algebras.) The (tensor) category of modules with a derivation is the same as the category of modules over the formal group ring H, so holomorphic vertex algebras are the same as the commutative ring objects in this category.

What is the difference between a commutative algebra over $\hat{G}_a$ and a (non holomorphic) vertex algebra? The only difference is that expressions like $a^x b^y c$ are no longer holomorphic in x, y but can have singularities; more precisely $a^x b^y c$ lies in $V[[x,y]][x^{-1}, y^{-1}, (x-y)^{-1}]$. In other words we can provisionally define a vertex algebra to be a module V such that we are given functions $a^x b^y c^z \cdots$ for each $a, b, c, \ldots \in V$ which behave just like the corresponding functions for commutative rings over $\hat{G}_a$, except that they are allowed to have certain sorts of singularities. Notice that we can no longer reconstruct a commutative ring structure on V by defining $ab = a^x b$ at $x = 1$, because $a^x b$ may have a singularity at $x = 1$.

The definition above is too vague to be useful, so we try to make it more precise. What we really want to do is to define some sort of category, whose multilinear maps are somehow allowed to have the sort of singularities above, and whose commutative ring objects are just vertex algebras. We first ask in what sort of categories we can define commutative ring objects. The obvious answer is tensor categories, such as the category of modules over $\hat{G}_a$ (or over any cocommutative Hopf algebra) but this turns out to be too restrictive. (We will implicitly assume that all categories are additive and have some sort of "symmetric" structure.) A tensor category requires that multilinear maps should be representable, but this is sometimes not the case for the categories we are interested in, and in any case this assumption is unnecessary. It is sufficient to assume that for each collection of objects $A_1, \ldots, A_n, B$ of the category we are given the space of multilinear maps $Multi(A_1, \ldots, A_n, B)$, and that these satisfy a large number of fairly obvious properties which I cannot be bothered to write down.

Unfortunately multilinear categories are not really the right objects either. The problem is the following. An expression like $a^x b^y c$ should live in a space like $V[[x,y]][x^{-1}, y^{-1}, (x-y)^{-1}]$. However the expression $a^x(b^y c)$ does not naturally live in this space, but in the larger space $V[[y]][y^{-1}][[x]][x^{-1}]$, and $(a^{xy^{-1}}b)^y c$ lives in a different larger space. This makes it hard to compare these expressions in a clean way. The easiest way to solve this problem is to define it out of existence by using "relaxed multilinear categories". The key idea is that instead of just once space of multilinear maps $Multi(A_1, \ldots, A_n, B)$ we are given many different spaces $Multi_p(A_1, \ldots, A_n, B)$ of multilinear maps, parameterized by trees p with a root (corresponding to B) and n leaves, corresponding to $A_1, \ldots, A_n$. We should also have some extra structures, consisting of maps between different spaces

of multilinear maps corresponding to collapsing maps between trees, and a composition of multilinear maps taking multilinear maps of types $p_1, \dots, p_n, p$ to a multilinear map of type $p(p_1, \dots, p_n)$. (Here $p(p_1, \dots, p_n)$ is the tree obtained by attaching $p_1, \dots, p_n$ to the leaves of p.) For details see my paper or Soibelman's notes, or better still work them out for yourself.

The main point is that in a relaxed multilinear category it is still possible to define commutative associative algebras. Joyal has pointed out that the definition of associative algebras in a relaxed multilinear category is strikingly similar to the definition of an A_∞ algebra; for example, the cells of the complexes used to define A_∞ algebras are parameterized by rooted trees with n leaves, and the boundary maps correspond to the collapsing maps between trees.

One way of constructing relaxed multilinear categories is as the representations of "vertex groups". A vertex group can be thought of informally as a group together with certain sorts of allowed singularities of functions on the group. More precisely a vertex group is given by a cocommutative Hopf algebra H, which we thing of as its group ring, together with an algebra of "singular functions" K over the "coordinate ring" H^* of H. The axioms for a vertex group say that K behaves as if it were the ring of meromorphic functions over the "group" G; for example, the ring of meromorphic functions is acted on by left and right translations, so K should have good left and right actions of H. A typical example of a vertex group is to take $H = k[D]$, $H^* = k[[x]]$ (so that H is the formal group ring of $\hat{G}_a$), and to take K to be the quotient field $k[[x]][x^{-1}]$ of H^*, which we can think of as the field of rational functions on the formal group $\hat{G}_a$.

We can construct a relaxed multilinear category from a vertex group roughly as follows. The underlying category is the same as that of the Hopf algebra of the vertex group. However the spaces of multilinear maps are different. Rather than define these in general, which is a bit complicated, we will just look at one example. We take G to be the vertex group above (whose commutative rings are vertex algebras), and take 3 G-modules A, B, and C. Then the space of bilinear maps from A, B to C is defined to be the ordinary space of bilinear maps from $A \times B$ to $C[[x, y]][(x - y)^{-1}]$ which are invariant under an action of G^3. (The easiest way to work out what the action of G^3 should be is to see what it has to be for the invariant bilinear maps from $A \times B$ to $C[[x, y]]$ taking $a \times b$ to $\sum_{i,j} f(D^i a, D^j b) x^i y^j / i! j!$ to be the same as invariant maps f from $A \times B$ to C.)

We summarize what we have done so far:

1 We have introduced "vertex groups"'.

2 The modules over a vertex group form a "relaxed multilinear category".

3 The commutative ring objects over the simplest nontrivial vertex group are exactly vertex algebras.

Now that we have set up this machinery, it is easy to find higher dimensional analogues of vertex algebras: all we have to do is look at commutative algebras over higher dimensional vertex groups G; we will call these vertex G algebras. As an example we will construct vertex algebras related to free quantum field theories in higher dimensions. (Can one construct vertex algebras corresponding to nontrivial quantum field theories in higher dimensions? At the moment this is just a daydream, as it is too vague to be called a conjecture.)

We first need to construct a suitable vertex group G. We take its underlying Hopf algebra H to be the polynomial algebra $\mathbf{R}[D_0, \ldots, D_n]$ where $D_i = \partial/\partial x_i$, which we think of as the universal enveloping algebra of the Lie algebra of translations of spacetime (with $x_0 = t$). The dual H^* is then $\mathbf{R}[[x_0, \ldots, x_n]]$, which we think of as the algebra of functions on spacetime. We define K to be $H^*[(x_0^2 - x_1^2 - \cdots x_n^2)^{-1}]$, which we think of as the algebra of functions on spacetime which are allowed to have singularities (poles) on the light cone.

Now we define the vertex G algebra V. The underlying space of V is the universal commutative H-algebra generated by an element ϕ, so $V = \mathbf{R}[\phi, D_0\phi, D_1\phi, \ldots, D_0^2\phi, \ldots]$ is a ring of polynomials in an infinite number of variables. We think of V as the ring of classical fields generated by ϕ, and it is a (holomorphic) vertex G algebra as it is a commutative ring acted on by H. We will turn it into a a nontrivial vertex G algebra by "deforming" this trivial vertex G algebra structure. (In general, for vertex G-algebras, quantization means deforming the structure on some commutative ring to turn it into a vertex G algebra.)

To do this we recall the following method of constructing commutative rings: if V is a space acted on by commuting operators v_n, and if V is generated by an element $1 \in V$ by the action of these operators, then V has a unique commutative ring structure such that 1 is the identity and the actions of all the operators are given by multiplication by elements of V. (Proof: easy exercise.) A similar theorem holds for vertex algebras (as was proved by Frenkel, Kac, Rado, and Wang). We will make V into a vertex G algebra by finding a vertex operator $\phi(x) = \phi(x_0 \ldots, x_n)$ acting on V such that $\phi(x)\phi(y) = \phi(y)\phi(x)$ and applying the construction above.

To construct $\phi(x)$, we first put

$$\phi^+(x) = \sum_i D^i \phi x^i / i!.$$

The vertex algebra structure on V defined by this vertex operator is just the commutative ring structure on V, so we need to deform ϕ^+. We define $\phi^-(x)$ to be the unique G-invariant derivation from V to $V[x][(x_0^2 - x_1^2 \cdots)^{-1}]$ taking ϕ to some even function $\Delta(x)$ (called the propagator). This is uniquely defined by the universal property of V. Finally we put

$$\phi(x) = \phi^+(x) + \phi^-(x).$$

It is easy to check that $\phi(x)$ and $\phi(y)$ commute, as $\phi^+(x)$ and $\phi^+(y)$ commute, $\phi^-(x)$ and $\phi^-(y)$ commute, and $[\phi^-(x), \phi^+(y)] = -[\phi^+(x), \phi^-(y)] = \Delta(x - y)$. Therefore we can make V into a commutative vertex G-algebra.

Notice that in quantum field theory, $\phi(x)$ means the value of some operator valued distribution ϕ at some point x of a manifold. On the other hand, for vertex G algebras, $\phi(x)$ should be thought of as the action of a "group" element x on an element ϕ of a vertex G algebra. Several other concepts in quantum field theory can also be translated into vertex algebra theory; for example, the correlation functions are $Tr(\phi(x)\phi(y)\phi(z)\cdots)$, where Tr is some G invariant linear function on V. Some concepts are not so easy to extend; for example, vertex G algebras do not seem to be able to cope with arbitrary curved spacetimes other than Lie groups, and it can be difficult to reconstruct a Hilbert

space (this includes as a very special case the problem of deciding which representations of the Virasoro algebra are unitary).

Finally we will briefly describe a special case of a theorem about vertex G-algebras generalizing the usual identity for vertex algebras. This theorem says (roughly) that under fairly general conditions, the integral of a vertex operator over an n-cycle is a vertex differential operator of order n. We will illustrate this for ordinary vertex algebras, when it says that the integral of $a(x)$ around the origin is a vertex differential operator of order 1. This means

$$\int a(x)dxb(y) - b(y)\int a(x)dx$$

has order 0 and is therefore of the form $t(y)$ for some t. Applying both sides to 1 and taking $y = 0$ shows that $t = \int a(x)dxb$. Therefore

$$\int a(x)dxb(y)c - b(y)\int a(x)dxc = (\int a(x)dxb)(y)c.$$

Finally integrating both sides around $y = 0$ shows that

$$\int a(x)dx \int b(y)dyc - \int b(y)dy \int a(x)dxc = \int (\int a(x)dxb)(y)dyc.$$

which is a special case of the vertex algebra identity. (The other cases can be deduced in a similar way.)